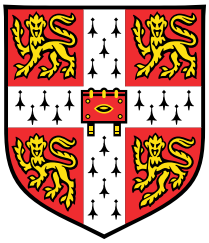
Integrative statistical methods for the genomic analysis of immune-mediated disease



Oliver S Burren

Department of Medicine
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College June 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Oliver S Burren
June 2019

Abstract

Genome wide association studies (GWAS) have proved to be a successful method in cataloguing loci influencing thousands of complex human disease phenotypes. However, elucidating the causal mechanisms underlying such associations has proved challenging due to the regulatory nature of the majority of signals.

In Chapters 2 and 3, I hypothesised that promoter-capture Hi-C (PCHi-C) data might have utility in physically linking disease-associated regulatory variants to their target genes, in a tissue-specific manner. To examine the genome-wide enrichment of GWAS summary statistics within PCHi-C chromatin contact maps I developed a novel statistical method, '*blockshifter*'. I applied *blockshifter* to a compendium of GWAS summary statistics for 31 traits and PCHi-C data across 17 primary blood tissues, and found convincing evidence for the enrichment of immune-mediated disease (IMD) GWAS signals in lymphocyte specific chromatin interactions, providing support for the hypothesis. Taking a more gene-centric approach I developed 'COGS', a novel method for integrating GWAS and PCHi-C to prioritise specific causal variants, genes and cellular contexts for functional follow up. With a focus on IMD, I prioritised tissue-context specific interactions in $CD4^+$ T cells linking putative causal variants for type 1 diabetes, to the promoter of *IL2RA*. The effect of these variants on *IL2RA* expression was subsequently validated by allele specific expression, by a collaborator, supporting the approach.

In Chapter 4, I hypothesised that summary statistics from multiple, well powered GWAS of related diseases might be exploited to provide insight into rarer related diseases or disease subtypes. To investigate this I developed a PCA based framework to generate a lower dimensional basis, summarising input GWAS traits. I constructed such a basis from ten IMD GWAS studies, excluding variants in the HLA region, and projected on summary GWAS data from multiple sources in order to characterise individual principal components (PCs). By projecting on both summary and individual level genotype data for juvenile idiopathic disease subtypes, I was able to show that a single PC was able to discriminate enthesitis-related and systemic forms of the disease from other subtypes.

I would like to dedicate this thesis to my wife Amanda.

Acknowledgements

Firstly I extend my thanks to my supervisor, Dr. Chris Wallace, without whom this thesis would not have been possible for many reasons. Thanks also to my co-supervisor, Dr. Mikhail Spivakov, for helping me navigate the history of chromatin dynamics and gene regulatory processes. I owe a debt of gratitude to the denizens of The Fishbowl, past and present, Dr. Mary Fortune, Dr. Jonny Griffiths, Mr. Martin Kelemen, Ms Kath Nicholls and Mr. Stephen Coleman. To Dr. Stasia Grinberg and Dr. James Liley, my longer term colleagues, thanks for your companionship, help with mathematics and its notation, and general support over the past three years.

From my previous laboratory, I would like to take the opportunity to say thanks to Prof. John Todd and Prof. Linda Wicker, who afforded me multiple opportunities for both personal and professional development during my tenure at the Diabetes and Inflammation Laboratory (DIL). Thanks are also due to Dr. Tony Cutler and (soon to be Dr.) Dan Rainbow, whose immunological and empirical expertise, generously given, stimulated my interest in this area. Thanks also to Prof. Ken Smith for taking a chance on me and introducing me to rarer forms of immune-mediated disease.

I have been lucky to have had input from a number of clinicians, over the course of this work but would like to especially thank, Prof. Lucy Wedderburn for her clinical insights into JIA and Dr. Jagtar Singh for the discussions, on treatments, and the clinical course of more adult forms of arthritides.

My deepest gratitude is to my family; to my wife, Mands for both the emotional and practical side of supporting a post-graduate husband, and my children, Alice and Eve, for keeping me tethered (my ‘book’ is now finished!).

Finally I would like to acknowledge all the patients, healthy controls and researchers on whose samples and analyses this thesis would not have been possible.

Table of contents

List of figures

List of tables

List of Abbreviations

Acronyms / Abbreviations

3C Chromosome conformation capture

4C Chromosome conformation capture-on-chip

5C Carbon-copy chromosome conformation capture

aBF Wakefield's asymptotic Bayes Factor

ATAC-Seq Assay for transposable accessible chromatin using sequencing

CHi-C Capture Hi-C

ChIA-PET Chromatin interaction analysis by paired-end tag sequencing

ChIP-Seq Chromatin immunoprecipitiation and sequencing

COGS Capture Hi-C omnibus gene score

cSNP Protein-coding single nucleotide polymorphism

CV Coefficient of variation

DHS DNase I hypersensitivity site

eQTL Expression quantitative trait locus/loci

eRNA Enhancer RNA

EUR The 1000 genomes European population cohort

GO Gene ontology

GSEA Gene-set enrichment analysis

GWAS Genome-wide association study

HCC Hepatocellular carcinoma

HLA Human leukocyte antigen

HPO Human phenotype ontology

HWE Hardy-Weinberg equilibrium

IBD Inflammatory bowel disease

ILAR International League of Associations for Rheumatology

IMD Immune-mediated disease

LCR Locus control region

LD Linkage disequilibrium

MAF Minor allele frequency

mCOGS Multiple casual variant COGS

miRNA Micro-RNA

MPRA Massively parallel reporter assays

mRNA messenger RNA

MVN Multivariant normal distribution

NCP Non-centrality parameter

OLS Ordinary least squares

OR Odds ratio

pBF Pseudo-Bayes factor

PCA Principal component analysis

PCHi-C Promoter capture Hi-C

PIR Promoter interacting region

PMI Poor man's imputation

PPD Preaxial polydactyly

PRS Polygenic risk score

sCOGS Single casual variant COGS

sCVPP Single causal variant posterior probability

snoRNA Small nucleolar RNA

SNP Single nucleotide polymorphism

snRNA Small nuclear RNA

SRD Self-reported disease

SVD Singular value decomposition

TAD Topologically assocaited domains

TSS Transcriptional start site

UKBB United Kingdom biobank

VPF Virtual promoter fragment

WTCCC Wellcome Trust Control Consortium

Chapter 1

Introduction

1.1 Foreword

The main aims of this thesis are threefold:

Aim One: Investigate whether the promoter interacting regions (PIRs) identified by promoter-capture Hi-C (PCHi-C) are enriched for GWAS signals in a cell-context specific manner.

Aim Two: Develop methods to integrate GWAS signals with PCHi-C data in order to prioritise putatively causal SNPs, genes and tissue contexts for functional followup.

Aim Three: Develop a framework for constructing a summary of the genetic relationships between multiple immune-mediated diseases (IMD) and evaluating how they might be useful in characterising rarer or clinically heterogeneous IMDs.

The work presented in this thesis is, like a majority of contemporary research, of a cross-disciplinary nature encompassing the fields of genetics, genomics and statistics. Given this scope, I have organised the material such that each subsequent chapter contains a more specific introduction to the relevant concepts, studies and literature that it is concerned with. In contrast, this introductory material is of a more general nature covering key concepts and technologies that form a foundation for subsequent chapters. In Section 1.2, to provide background to aims one and two, I describe how eukaryotic genomes are organised, the main empirical methods for measuring facets of this organisation (including PCHi-C), and how such genomic organisation is of relevance to human disease. Section 1.3

introduces key population genetic concepts and statistical frameworks that I rely on throughout this thesis to achieve my aims. In the final section (Section 1.4), I provide a general introduction to IMDs with an emphasis, relating to aim three, on their shared and distinct genetic architectures. I finish by briefly touching on how the integration of genetic and functional data might afford a more, robust, molecular taxonomy of IMDs.

1.2 The role of genome organisation in health and disease

In 1958 'On protein synthesis' was published setting out Sir Francis Crick's 'Central Dogma' on how information stored in DNA could give rise to the complex biochemistry essential for life (Crick, 1958). In it, he stated a flow of information from DNA, which through transcription to intermediate RNA species, is ultimately, translated to proteins, in order to elicit cellular function.

Five years later Monod and Jaques were the first to characterise this process in prokaryotes using the polycistronic *lac* operon in *Escherichia coli* (Jacob and Monod, 1961). Empirically, they demonstrated the presence of 'regulatory' genes and sequence elements, whose function was to control the activity of a set of target genes. These regulatory genes, which we now call 'transcription factors', were shown to function by interacting with the cognate DNA sequence elements to regulate the expression of a short lived intermediate that they called 'messenger' RNA (mRNA).

1.2.1 The canonical eukaryotic protein coding gene

At the sequence level, the canonical eukaryotic protein-coding gene, which I refer to subsequently as a 'gene', consists of multiple elements that are required for functional transcription.

The Promoter Found directly upstream of the transcriptional start site (TSS), the promoter initiates binding of RNA Polymerase II (Pol II), the enzyme responsible for transcribing DNA to mRNA. Generally such promoter sequences consist of two elements; a region immediately upstream of the TSS known as the 'core promoter' and a region upstream to this, known as the 'proximal element' or 'regulatory promoter' (Kanhere and Bansal, 2005). The former provides sequence cues for the binding of the Poll II complex,

with the latter thought to provide a more subtle, context-specific modulation of expression rate through the binding of cofactors such as transcription factors. As a result promoters are highly heterogeneous between genes reflecting different abilities to drive transcription in different tissue contexts.

The Enhancer In Eukaryotes the activity of the promoter in controlling transcription is augmented by actions of short (between 100-500 bp) sequences, known as enhancers. Enhancers function by the binding of specific transcription factors that once recruited interact with co-factors and Pol II to potentiate transcription of a target or set of target genes. In higher organisms, enhancers, unlike promoters are often found at some distance (up to 1Mb) either upstream or downstream from their target gene. This effect over distance means that they are often found in the intronic regions of non-target genes or even 'skip' multiple intervening genes to exert their function.

To understand such action at a distance it is useful to summarise current knowledge about the organisation of DNA within a eukaryotic cell.

1.2.2 Chromatin structure

In humans, a majority of cells contain a complete copy of an individuals genetic material. At a large scale this can be observed as classical karyotype consisting of 22 pairs of autosomal chromosomes and one pair of sex determining allosomes. This large body of genetic material must be efficiently packed into the cell nucleus, a specialised sub-cellular organelle. At the lowest level, DNA polymers associate with specialised proteins called histones that spool the DNA into approximately 146 bp supercoils known as nucleosomes (Figure 1.1a) to form euchromatin (Figure 1.1b) (Higgs et al., 2007). In turn euchromatin can be further compacted to form the higher order structure of heterochromatin (Figure 1.1c and d). Generally euchromatin is an indicator for more active chromosomal regions whereas heterochromatin identifies those that are more quiescent. Indeed, The formation of euchromatin and heterochromatin is highly regulated (Lawrence et al., 2016) through multiple mechanisms.

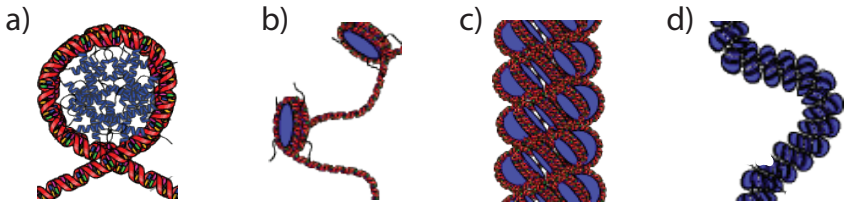


Fig. 1.1 A model for the formation of chromatin in Eukayotes. **a)** Specialised proteins called core histones (blue) form ≈ 146 bp DNA coils called nucleosomes. **b)** Nucleosomes form at regular intervals along the DNA double helix in the so called 'beads on a string' configuration. This configuration is permissive to gene transcription and is modulated by chemical modification of histone tails. **c)** Mediated by the non core histone, H1, this 'bead on a string' configuration, known as euchromatin, is further packed into fibres known as chromatin. **d)** These densely packed 30nm chromatin fibres, known as heterochromatin are generally less permissive to gene transcription. *Image adapted from an original image by Richard Wheeler under CC BY-SA 3.0 license* [https://commons.wikimedia.org/wiki/File:Chromatin_Structures.png].

1.2.3 Regulation of chromatin state

DNA-Methylation

At the DNA level, the addition of methyl groups to individual cytosine (C) bases is widespread. In mammals this methylation occurs at specific di-nucleotides known as CpG's (5'-C-phosphate-G-3'). This methylation is context specifc, for example, the methylation of gene promoter regions is associated with attenuation of gene transcription, where as methylation of gene bodies has a reciprocal relationship (Jones, 2012). It is thought that DNA-methylation indirectly affects chromatin structure by recruiting, methyl-CpG-binding domain proteins (MBDs) which with co-factors promote chromatin remodelling (Du et al., 2015) .

Covalent histone modification

At the unit of the nucleosome, constituent histone proteins have polypeptide 'tails' (Figure 1.1b) that through the action of specific enzymes may be covalently modified. For example, Histone 3 (H3) has a specific lyseine at position 27 (K27) that when acetylated (ac) correlates with more active chromatin (H3K27ac). Many such histone modifications have been described, correlating with different chromatin activation levels which are reviewed in (Lawrence et al., 2016).

Key to the study of histone modifications has been the development of ChIP-seq, which has made it possible to catalogue the location of specific modifications on a genome-wide scale across multiple tissues and organisms. ChIP-seq is an umbrella term for the process of using antibodies raised to specific DNA binding proteins (Landt et al., 2012). These antibodies are extremely specific and therefore can be used to precipitate histones with particular covalent modifications in complex with the DNA to which they associate. This DNA can then be sequenced and mapped back to the genome, allowing the elucidation of the physical location of the modified nucleosome. The ChIP-seq method can be applied to any DNA interacting protein for which a specific antibody can be raised, and has been used successfully to interrogate and characterise the binding of many transcription factors (The ENCODE Project Consortium, 2012).

The mechanisms by which such modifications are able to affect chromatin state are varied, for example, H4K16ac is thought to loosen intra-nucleosome binding, thus favouring the formation of euchromatin and transcription factor accessibility (Shogren-Knaak et al., 2006).

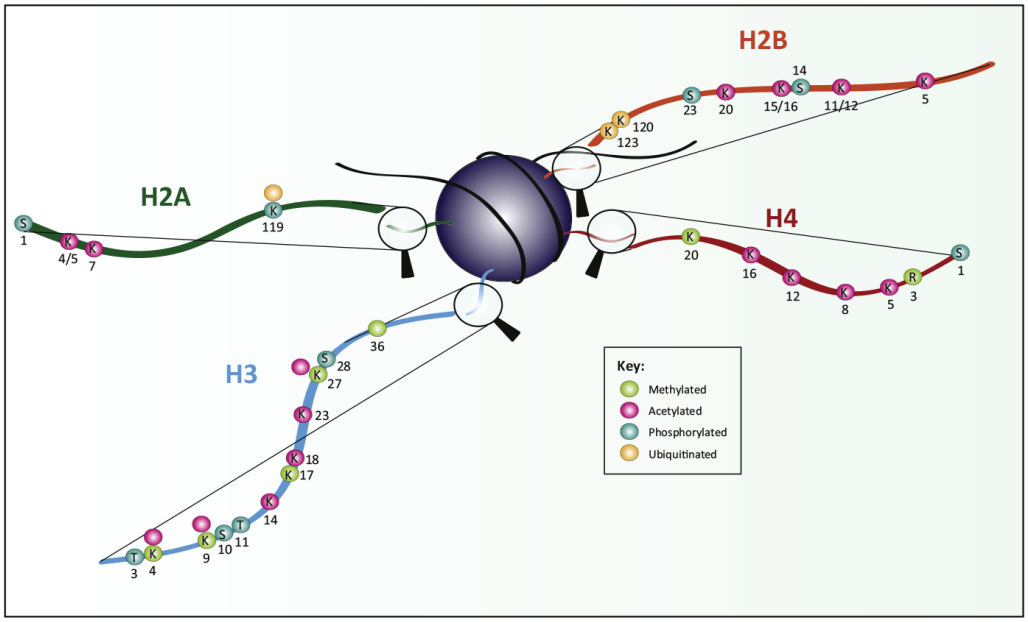


Fig. 1.2 A catalogue of covalent histone modifications in eukaryotes. The central sphere indicates a nucleosome complexed with DNA (black). Histone tails are blown up for clarity with specific residues and their possible modifications numbered. *Image adapted from an original from Lawrence et al. (2016), ©Elsevier Ltd. reproduced under license.*

DNA accessibility

The accessibility of chromatin and its constituent DNA is also a useful marker of activity. One such assay, DNase hypersensitivity (DHS) assay with sequencing (DNase-Seq), involves digesting chromatin with the DNA cleaving enzyme DNase I (Song and Crawford, 2010). Regions of open chromatin are more accessible to the enzyme, and are therefore more likely to be a cleaved. These cleavage products, can be isolated and sequenced and their physical locations, known as DHS regions, found by mapping to a reference genome. The DHS assay is technically challenging and more recently has been replaced with the assay for transposase-accessible chromatin using sequencing (ATAC-Seq) which requires less biological input material (Buenrostro et al., 2015).

Chromatin segmentation

Due to the number of modifications that individual nucleosomes can undergo the number of possible combinations is large (Figure1.2), requiring the combined analysis of heterogeneous sources of data derived from CHiP-Seq, ATAC-Seq and RNA-Seq experiments. This precipitated the development of software to integrate datasets performed on the same cell type to identify patterns of modifications, transcription factor binding, accessibility and transcription that correlate with chromatin activity (Ernst and Kellis, 2012; Hoffman et al., 2012). This has allowed the annotation of tissue specific 'chromatin segments', regions of chromatin with similar properties (e.g. combinations of histone modifications) (Ernst and Kellis, 2015) and their subsequent characterisation into more conceptual constituent elements such as enhancers.

1.2.4 Chromatin organisation in three dimensions

However, this linear cataloguing of the non-coding portion of the genome described in the previous section, misses additional complexity, in that chromatin fibres associate to form higher order three dimensional structures. Whilst some of these associations are structural, allowing the very long fibres of chromatin to be efficiently packed within a cell, many have specific functions associated with replication and the regulation of gene expression. The recent development of high-throughput methods, known as chromatin conformation assays, for assessing this three dimensional structure has begun to reveal the underlying complexity of this structure and how it effects cellular function.

Chromatin conformation assays

Building on the Nuclear Ligation assay (Cullen et al., 1993) Dekker et al. (2002) described a molecular technique, chromosome conformation capture (3C) for interrogating intact nuclei for the presence of specific interactions (Figure 1.3). Subsequently this was extended through, chromosome conformation capture-on-chip (4C), to examine all interactions *with* a specific locus (Simonis et al., 2006). Further development gave rise to carbon copy chromosome conformation capture (5C), that allowed the interrogation of all interactions *within* a specific locus of up to 1Mb in size (Dostie et al., 2006). These techniques for investigating higher order chromatin structure dovetailed with the emergence of massively parallel sequencing techniques leading to the development of Hi-C.

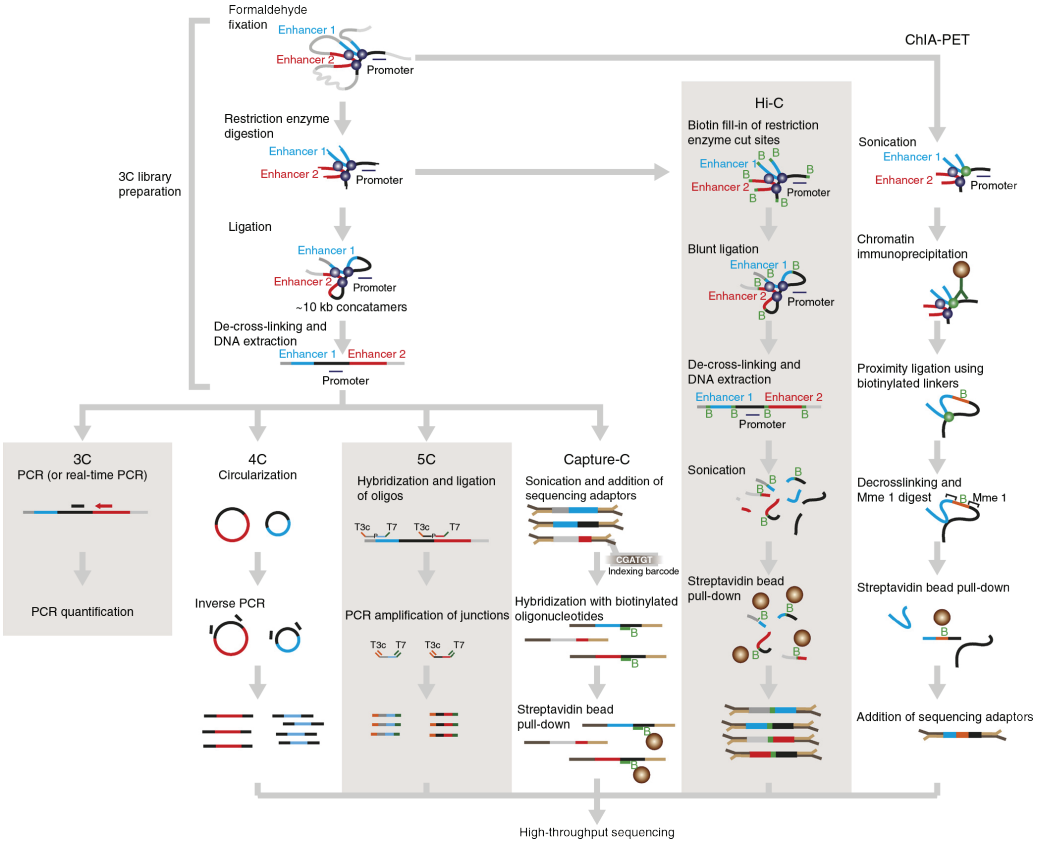


Fig. 1.3 Methods for interrogating chromatin conformation. *Image adapted from an original from Davies et al. (2017), ©Springer Nature reproduced under license.*

Hi-C

Hi-C involves cross-linking genomic DNA with formaldehyde resulting in covalent links between spatially adjacent chromatin segments. This chromatin is then

digested with a restriction enzyme and sticky ends are filled in with biotin labelled nucleotides. Ligation is then performed under dilute conditions, thus favouring intra-molecular ligation events. The DNA is then purified and then sonically sheared and fragments are then enriched for biotinylated junctions, which then undergo paired end sequencing (Lieberman-Aiden et al., 2009) (Figure 1.3).

Hi-C resolution is limited by two main factors. Firstly, the protocol involves a restriction enzyme digest, usually *Hin*dIII, and interactions are called based on the resultant fragments generated, in practice if *Hin*dIII is used this limits resolution to approximately 4Kb. Secondly, the complexity of the sequence libraries generated means that to increase the effective resolution by a factor n requires an n^2 fold increase sequencing reads which is prohibitive for general implementation (Jäger et al., 2015).

ChIA-PET and Capture-C

In a parallel, an alternative experimental method, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) was being developed, that harnessed immunoprecipitation techniques to enrich for interactions based on the presence of specific DNA binding proteins (Fullwood et al., 2009). The main limitation of ChIA-PET is that the ChIP step results in a relatively modest enrichment for targeted chromatin interactions. Consequently the number of sequence reads available for mapping genuine contacts is attenuated resulting in reduction in the sensitivity to uncover interactions (Davies et al., 2017). More recently the development of hi-ChIP (Mumbach et al., 2016) has overcome some of these limitations, resulting in meaningful reductions in the amount of biological input material required. However, both ChIA-PET and hi-ChIP still rely on specific antibody binding to target proteins, which themselves are subject to bias in their efficacy (Davies et al., 2017).

In contrast to the immunoprecipitative methods mentioned previously, other chromatin methods use sequence based DNA capture techniques to enrich for contacts containing specific sequences of interest. Capture-C (Hughes et al., 2014) combines 3C techniques with sequence specific capture allowing a detailed picture of chromatin contacts at 100's of individual 'viewports' to be assembled (Davies et al., 2016). Whilst capture-C overcomes the main limitations of other techniques (e.g. high biological material input requirements and reliance on protein specific iummunoprecipitation) it is limited to the one-to-many contact nature of 3C and

subsequent multiplexing considerations limit the number of viewports that can be assayed to the hundreds.

Capture Hi-C

These limitations have lead to the parallel development of Capture Hi-C (CHi-C) (Dryden et al., 2014). This method combines *in situ* ligation adaptations to conventional Hi-C methodology using a sequence capture library design to target specific regions of the genome. The promoter capture Hi-C (PCHi-C) methodology used extensively in this thesis, for example, targets the 5' and 3' of restriction fragments that overlap gene promoters genome-wide (Mifsud et al., 2015). This results in the subsequent enrichment of distal sequences that come into contact with the promoter regions targeted by the capture design.

It should be stressed that whilst some of the techniques described above have been subsumed by subsequent developments, there is no 'best' technique. At the genome-wide scale Hi-C can give a relatively non-biased overview of 3D genome topology useful for the understanding of how higher level chromatin organisation operates genome-wide. In contrast capture-C can be used to characterise specific interactions identified through orthologous empirical methods (e.g. ChIP-Seq or ATAC-Seq) in detail. As such PCHi-C occupies the middle ground operating as a bridge between genome-wide Hi-C and locus specific capture-C, important for transducing information between the two scales of chromatin organisation.

Chromatin looping

An important locus for deriving tissue specific mechanisms of enhancer action over a distance has been the murine locus control region (LCR) responsible for the expression of the $\beta-$globin gene cluster. The LCR itself contains multiple tissue context specific enhancers, the most distal of which is located 70kb from the $\beta-$globin gene locus. Tolhuis et al. (2002) were able to show, using 3C, that the tissue specific expression of the α and $\beta-$globin genes were modulated by the physical interaction of specific LCR enhancers to the promoters of these genes, looping out the intervening $\varepsilon\gamma$ and βh1 genes.

Topologically associated domains (TADs)

At a more global level Dixon et al. (2012) used Hi-C to describe the phenomenon of topologically associated domains(TADs). These domains are areas of chromatin

that interact with increased frequency when compared with those outside of the TAD, and their boundaries are enriched for the DNA binding proteins CTCF and cohesin. Chromatin looping events identified through Hi-C were found to also preferentially occur within rather than between TADs (Dixon et al., 2012; Rao et al., 2014; Sexton et al., 2012), an observation that was subsequently validated using capture-C (Hughes et al., 2014). Whilst such TAD architecture seems to be independent of tissue context (Dixon et al., 2012) its ablation, through the removal CTCF binding domains can have significant localised effects on the regulation of gene expression (Zuin et al., 2014). Recent work in mice has demonstrated that novel tissue-specific TAD borders can occur at promoters of developmentally regulated genes. These borders can be separated by the differential enrichment of DNA-specific binding proteins such as cohesin and histone modifications such as H3K4me3 and H3K27ac (Bonev et al., 2017). The question remains as to whether the correlation between such epigenetic marks and the regulation of chromatin organisation is causative. Whilst this is an active area of research, recent work has shown provisional support for H3K4me1 having a causal role in the stabilisation of long range chromatin looping events through the active recruitment of cohesin (Yan et al., 2018).

1.2.5 Chromatin organisation and disease

In the previous section I discussed the role and main mechanisms underlying the choreography between the 3D genome and gene expression. An outstanding question is to what extent alterations at the sequence level can attenuate or ablate chromatin contacts and thereby alter tissue specific transcriptional programmes to cause disease?

In rare monogenic disease

An early example of how ectopic long range chromatin looping could be modulated through the actions of a single nucleotide polymorphism (SNP) was observed through the genetic mapping of preaxial polydactyly (PPD). Using the *sasquatch* (*Ssq*) mouse model for PPD, Lettice et al. (2003) characterised a prominent limb enhancer (ZRS) modulating *Shh* gene expression over 1Mb away responsible for the PPD phenotype. Using segregation analysis in multiple human families exhibiting the PPD phenotype they were able to show that all affected individuals were homozygous for SNPs overlapping the human-syntenic region of ZRS. In this highly penetrant monogenic setting, where homozygosity for a specific and

rare non-coding allele segregates with disease status, there are few additional examples. Indeed, whilst sequencing studies have identified a large volume of protein coding SNPs responsible for mongenic disease (Lek et al., 2016), a recent study investigating the effect of rare variation on neurodevelopmental disorders estimated that between 1-3% of cases might be caused by *de novo* non-coding variation (Short et al., 2018).

In common polygenic disease

This situation is somewhat reversed in the context of common polygenic disease, where a majority of associated variants, identified through genome-wide association studies (GWAS) have been found to occur outside of the regions of the genome coding for proteins (1000 Genomes Project Consortium et al., 2015). This has been problematic because GWAS have not typically led to the identification of disease causing genes. This observation (which I expand upon in Section 1.4.4), alongside results emerging from model organisms such as yeast (Bloom et al., 2013), opened up the possibility that knowledge about 3D genome organisation and its interplay with gene regulation might have utility in the interpretation of causal mechanisms underlying complex traits. One such early study used 3C evidence to show that a Type 1 diabetes (T1D) association signal on chromosome 16, located within the intron of the gene *CLEC16A*, might instead regulate a more distal gene *DEXI* for which little biology was known (Davison et al., 2012). This study highlighted that the physical location of an association might be an imperfect method of prioritising genes for functional characterisation and that chromatin conformation capture might provide a valuable orthogonal method.

This was brought in to sharp contrast by Smemo et al. (2014), who investigated an obesity associated locus within the intron of the *FTO* gene on chromosome 16 (Frayling et al., 2007). Previous efforts to characterise the region, including mouse knockout studies, had provided support for causality of the *FTO* gene (Church et al., 2009). Using 4C Smemo et al. (2014) showed evidence for the interaction in mouse brain tissue of a region, syntenic for the human associated region, with the promoter of *Irx3* rather than *Fto* precipitating considerable controversy. This controversy was resolved by Claussnitzer et al. (2015) who used multiple sources of genomic evidence (including Hi-C) to elucidate a role for the human causal variant in abrogating *ARID5B* transcription binding. This in turn affected regulation of *IRX3* and *IRX5* expression through chromatin looping causing a downregulation of thermogenesis.

1.3 Statistical methods for genomic analysis

In the previous section I introduced the concept that common variants, through chromatin looping events, can modulate complex disease risk via the dysregulation of distal gene expression. Next I expand on the statistical techniques and challenges relevant to robustly identifying common causal variants in complex disease. My focus on methods for testing association rather than linkage is mostly technical as the high genetic heterogenity combined with low effect sizes underlying a majority complex disease genetic architectures considerably disadvantages the power of linkage compared to association based approaches (Risch and Merikangas, 1996).

1.3.1 Relevant population genetics concepts

Prior to discussing such statistical methods, it is worth touching on population genetics concepts that have particular relevance to the methods of statistical association I subsequently describe. In this section I introduce allele frequencies, how over many generations these vary and give rise to structure in genetic data. Finally I touch upon 'heritability' a concept that can be used to measure the relative contributions of environment and genetics to the variability of a phenotype within a population.

Allele frequency and Hardy-Weinberg equilibrium

In 1908, separate publications from Wilheim Weinberg and G. H. Hardy set out a mathematical framework for modelling allele frequencies within a population that would arise over many generations of random mating. Consider a SNP for a diploid organism consisting of a alleles A_0 and A_1. In a large population of size N the frequencies of A_0 and A_1 can be estimated as $p = \frac{C(A_0)}{2N}$ and $q = \frac{C(A_1)}{2N}$, $C(A_0)$ and $C(A_1)$ are the observed allele counts of A_0 and A_1 accordingly. In a large population p and q can be viewed as the probability of obtaining A_0 or A_1 from random sampling, and given that there are only two possible alleles $p+q=1$. An assumption of this model is that whilst each individual contains two alleles (as they are diploid), these are independently sampled from the population and so may be considered separately. A natural extension of this model is to consider the probability of obtaining a specific set of alleles or genotype when sampling an individual from the population. Given the previous observation of the probability of sampling a single allele (i.e. $p+q=1$) it follows that $(p+q)^2 = 1^2$. Here p^2

and q^2 correspond to the frequencies of the homozygous genotypes, A_0A_0 and A_1A_1 respectively and frequency of the heterozygous genotype, that exists in two configurations A_0A_1 and A_1A_0 is $2\sqrt{p^2q^2}$. This Hardy-Weinberg principle of equilibrium (HWE) relies on the assumption of stable allele frequencies within a population, and thus makes a number of implicit assumptions that relate to this, that include random mating, no population migration and that the target alleles are not under selection.

Linkage Disequilibrium

Due to the mechanism of meiosis, the process of recombination randomly shuffling genetic material between parental gametes, alleles at neighbouring SNPs become less correlated. On an individual level this results in a set of SNPs (normally spatially proximal) at different loci being non-randomly associated, these alleles are said to be in linkage disequilibrium (LD). To characterise LD pairwise between two SNPs various metrics have been defined (Devlin and Risch, 1995). By far the most commonly applied is the correlation coefficient r. Let P_A and P_B be the estimated minor allele frequency at SNPs A and B respectively where P_{AB} is frequency of the minor alleles at A and B co-occurring on the same chromosome then

$$r = \frac{P_{AB} - P_A P_B}{\sqrt{P_A(1-P_A)P_B(1-P_B)}}. \tag{1.1}$$

Here P_{AB} is equivalent to the haplotype[1] frequency of minor alleles at A and B. In general the square of the correlation coefficient is used in order to remove the sign introduced, as this arbitrarily dependent on the way in which alleles are labelled.

Heritability

Total heritability, H^2, is the proportion of the phenotypic variance, σ_P^2 that can be attributed to genetic differences, σ_G^2, among individuals (Equation 1.2).

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}. \tag{1.2}$$

This definition is flexible, for example, we might imagine σ_G^2 combining with an environmental factor, E, with a phenotypic variance σ_E^2 such that $\sigma_P^2 = \sigma_G^2 + \sigma_E^2$. However, this flexibility makes H^2 hard to estimate without making strong

[1] A group of alleles from the same chromosome

assumptions. Instead we might think of σ^2_G as the sum of variances across a range of additive (A), dominant (D) and interaction (I) effects, such that $\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_I$. This leads to the definition of narrow-sense heritability, which is the proportion of phenotypic variance explained by additive genetic effects

$$h^2 = \frac{\sigma^2_A}{\sigma^2_P}. \tag{1.3}$$

We can define a further quantity, h^2_g, called SNP heritability, as the proportion of variation in the trait that can be explained by additive effects of commonly-occurring SNPs whose genotype we can measure[2]. In practice $h^2_g \leq h^2 \leq H^2$ reflecting the increasing flexibility of their underlying definitions.

1.3.2 Genome Wide Association Studies

GWAS, employ, in a hypothesis-free manner, the methods described in the subsequent sections, to a phenotyped population for which the majority of common variation (MAF>5%) has been measured with the aim of uncovering whether such variation is associated with the target phenotype. These associations, once discovered, can implicate novel biological mechanisms underpinning the measured phenotype, for example autophagy in Crohn's disease (Zhang et al., 2008), and more recently are being used to stratify individual disease risk (Wray et al., 2019), a pre-requisite of 'personalised' medicine (Jameson and Longo, 2015). Whilst early studies examined the theoretical underpinnings of such approaches under a variety of scenarios (Wang et al., 2005), robust technologies to measure such a large amount of genotypes across a suitably powered cohort were in their infancy. The first reported GWAS, concerning age-related macular degeneration (Klein et al., 2005), assayed 160,000 SNPs across 96 cases and 50 controls, finding an association with the *CFH* gene, implicating a role for the complement system of innate immunity in disease pathogenesis. A watershed moment occurred in 2007 on the publication of the Wellcome Trust Case Control Consortium paper (Wellcome Trust Case Control Consortium, 2007). This married the technological breakthrough in the large scale measurement of SNP genotypes, begun with the international HapMap project (International HapMap Consortium et al., 2007), with a collaborative approach to data sharing and analysis enabling the first large scale GWAS, that included 2,000 cases for each of seven diseases and 3,000 controls,

[2] indeed this is also known by the pseudonym 'chip' heritability in reference to the underlying genotyping platforms employed

typed over 500,000 SNPs. Over the intervening years, iterative improvements to SNP genotyping platforms, statistical imputation methods and the aggregation of ever larger sample cohorts across a plethora of traits and diseases, has resulted in an exponential growth in the catalogue of human variation associated with human disease (Figure 1.4). For example a recent study of educational attainment assessed 1.1 million individuals across 9 million SNPs (Lee et al., 2018).

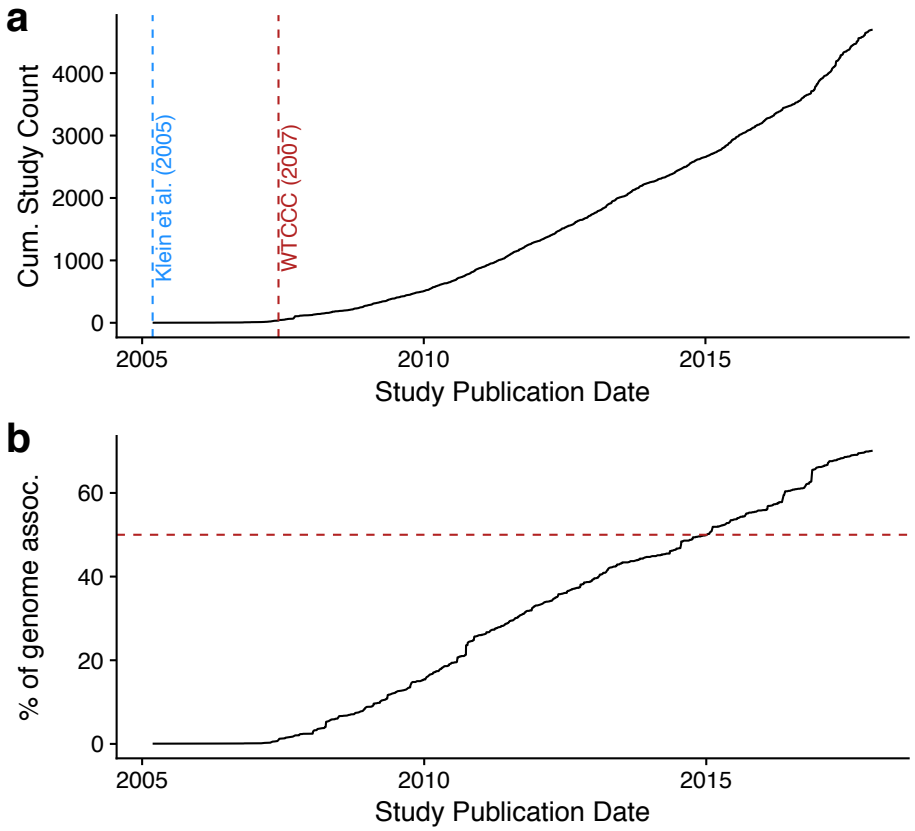


Fig. 1.4 Growth of Genome-Wide Association studies. **a)** Growth in the cumulative number of GWAS studies. For orientation Klein et al. (2005) and Wellcome Trust Case Control Consortium (2007) studies are shown. **b)** The cumulative growth of recombination blocks associated with one or more traits as a function of genome coverage over the same period. The red dotted line shows where half of the genome has been associated with one or more traits. Data downloaded from GWAS catalogue (25/01/2018).

1.3.3 Hypothesis testing

Hypothesis testing underlies many statistical inference approaches to GWAS. Key to the approach is defining a null hypothesis, H_0, that contradicts a theory about a particular phenomenon for which data has been collected. For example, consider

a random variable X with an observed value x, if the distribution of X under H_0 is known, $H_0 : X \sim f_X$, then it is possible to associate a probability of observing a value sampled from f_X by

$$\Pr(X = x \mid H_0). \tag{1.4}$$

More generally we might take repeated samples from f_X, selecting a statistic, known as a 'test' statistic, in order to provide a summary of those repeated samples. In most contexts this test statistic is devised such that it has a known distribution under H_0. Under the Neyman-Pearson approach, we partition this known H_0 distribution, into values of the test statistic which are unlikely to be observed if repeated samples of x are truly drawn from this distribution. This partition is called the critical region, and its probability is defined as α. If the sample test statistic, falls within this critical region, then H_0 is rejected and the alternative, H_1 is accepted. The value of α, is also known as the Type 1 error rate as it equates to the probability of erroneously rejecting H_0 when it should have been accepted. The Neyman-Pearson approach also defines β, the Type II error rate, which, in contrast, reflects the probability of erroneously accepting H_0, with it's value relating to both sample size and the nature of H_1. This approach to hypothesis testing is applied in a frequentist setting, where we assume that a parameter or test statistic of X has a true value, and that this is can be estimated by appropriate sampling from f_x. This is in contrast to a Bayesian philosophy, that assumes that the parameters are themselves random variables and as such have their own distributions. Thus, in the Bayesian setting, we instead estimate the *posterior* probability of H_0,

$$\Pr(H_0 \mid X = x) = \frac{\Pr(X = x \mid H_0)\ \Pr(H_0)}{\Pr(X = x)}, \tag{1.5}$$

which summarises how strongly H_0 is supported by the observed value, x. A key concept here is the requirement for selecting a prior probability for the null hypothesis, $\Pr(H_0)$ before observing x. This reflects our belief that H_0 is not fixed but is itself a random variable and as such has an associated probability of being true. In the field of genetics, frequentist approaches are used more widely, in contrast to Bayesian approaches, which are used for pragmatic reasons, such as for settings when sample size is small, there is a need to include prior information or where we wish to compare non-nested hypothesis (Section 1.3.5).

1.3.4 Frequentist approaches to genetic association testing

Generally, associations arise where we observe that one set of observations or events is statistically *dependent* on another. Most statistical tests for association utilise this dependence and examine the likelihood under the null hypothesis, H_0 that the two events are truly independent. It is important to note that such association tests in isolation cannot imply causality, a subject of debate outside the scope of this thesis. Tests of genetic association apply this principal of dependence to the field of genetics, by examining whether a quantitative or binary outcome is dependent on exposure to one or more underlying genetic variants.

In order to illustrate different approaches let us imagine a simplified fictitious study, where we measure the presence, A_1 or absence A_0 of a particular allele of a genetic variant, A in a set of N_1 individuals with disease, D^+, and N_0 without D^-. A natural method of summarising our findings is a 2×2 contingency table (Pearson and Blakeman, 1906) where each element reflects the frequency of two events.

	A_1	A_0
D^-	a	b
D^+	c	d

Table 1.1 A contingency of an association study where we wish to examine the statistical dependence of disease outcome (D^-/D^+) on genetic variant exposure (A_0/A_1)

Odds ratios

The main metric of effect size in binary contingency tables, the focus of our fictitious example, is the odds ratio (θ). To compute this we first compute the conditional probability of having the disease contingent on not having the variant ($Pr(D^+|A_0)$). We can express the odds of having the disease contingent on not having the variant as $\frac{Pr(D^+|A_0)}{1-Pr(D^+|A_0)}$, similarly the odds of having the disease contingent on having the variant is $\frac{Pr(D^+|A_1)}{1-Pr(D^+|A_1)}$. The odds ratio for disease associated with variant A_1 is thus:

$$\theta = \frac{\frac{\Pr(D^+|A_1)}{1-\Pr(D^+|A_1)}}{\frac{\Pr(D^+|A_0)}{1-\Pr(D^+|A_0)}}, \tag{1.6}$$

which using the labels from Table 1.1 simplifies to $\frac{cb}{ad}$.

This estimate of effect size does not tell us whether this effect is significant, such that it is meaningfully different from the value we would expect if D and A were independent (i.e. $\theta = 1$). In order for this we need to elucidate, given our study size, the uncertainty attached to our estimate of the odds ratio ($\hat{\theta}$). The distribution of $\hat{\theta}$ is skewed towards values greater than one because low values are constrained by zero whereas large values are not. This can be overcome by using the natural logarithm of $\hat{\theta}$ such that $\hat{\beta} = \log(\hat{\theta})$. We can compute an approximation of the standard error of $\hat{\beta}$ as $\sigma_{\hat{\beta}} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$, where a, b, c and d are the counts from the joint distributions of our contingency table (Table 1.1). Under the assumption that sample size is large, the estimate of our odds ratio, $\hat{\theta}$, follows a normal distribution, allowing us to obtain a standardised Z score ($Z = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$). This test statistic can then be used as the basis of a hypothesis test where under the null $\hat{\beta} = 0$, which is rejected if the Z-score intersects with the critical region defined by α.

Testing for association in biallelic single nucleotide polymorphisms

The previous section dealt with inference based on odds ratios in the case of two categorical variables. In general, most organisms are *ploidy* indicating that they have more than one set of chromosomes. Humans have a ploidy of two which means that for a given a binary allele where A_1 and A_0 indicate the allele is present or absent respectively, an individual is sampled from four possible configurations (Table 1.2). By updating our contingency table so the joint distributions reflect allele counts rather than individuals we can compute odds ratio's accordingly.

Configuration	Name	Value
A_0A_0	Homozygous A_0	0
A_1A_0	Heterozygous	1
A_0A_1	Heterozygous	1
A_1A_1	Homozygous A_1	2

Table 1.2 Value indicates the number of alleles present in a configuration

Linear Regression

In the previous section I discussed methods for estimating effect sizes and their significance given two sets of categorical variables. Here I introduce the linear model for testing for genetic association where the outcome variable is quantitative, rather than binary.

Let $\mathbf{g}$ be a vector of genotypes for a set of n individuals for a given biallelic SNP such that $\mathbf{g} = (g_1, g_2, \ldots, g_{n-1}, g_n), g_i \in \{0,1,2\}$ (Table 1.2). Let $\mathbf{y} = (y_1, y_2, \ldots, y_{n-1}, y_n)$. Generally we wish to understand whether $y_i = f(g_i) + \varepsilon$, where $f(g_i)$ is a linear function that can be parameterised by a slope, β and an intercept, β_0 such that $y_i = \beta_0 + \beta g_i + \varepsilon$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is an error term that allows for residual variability (i.e. that the value of y_i may be a function of other, unmeasured factors). We seek to optimise the selection of both β_0 and β such that any differences in the vector of predicted outcome variables, $\hat{\mathbf{y}}$, from applying the linear function of $\mathbf{g}$, are minimised with respect to the observed vector of outcomes, $\mathbf{y}$. We can estimate the difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$ as $\mathbf{y} - \hat{\mathbf{y}}$, which results in a vector of *residual* values, $\mathbf{r} = (r_1, r_2, \ldots, r_{n-1}, r_n)$ which can be summarised in the residual sum of squares quantity $\text{RSS} = \sum_i^n (r_i^2) = \sum_i^n (y_i - (\hat{\beta}_0 + \hat{\beta} g_i))^2$. There are multiple analytical procedures to solve for estimates of $\hat{\beta}_0$ and $\hat{\beta}_i$ that minimise RSS, however the ordinary least squares (OLS) method provides a closed form such that:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right)^T = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}, \tag{1.7}$$

where $\mathbf{X} = \begin{pmatrix} 1 & g_1 \\ \vdots & \vdots \\ 1 & g_n \end{pmatrix}$.

Such an approach can be extended to the genotypes of multiple SNPs and outcomes by replacing $\mathbf{g}$ with a genotype matrix $\mathbf{G}$, where rows and columns match samples and SNPs respectively and the elements of $\mathbf{G}$ belong to $\{0,1,2\}$ (Table 1.2).

Logistic Regression

Linear regression can be successfully used for inference when the outcome is quantitative, but it is less suitable when the outcome is binary such that $\mathbf{y} \in \{0,1\}$. Fitting a linear model is unsuitable as the model space allows for parameters that might result in predictions such that $\mathbf{y} < 0$ or $\mathbf{y} > 1$. An alternative approach is to employ a logistic function on $\mathbf{g}$ where:

$$f(g_i) = \frac{1}{1 + \exp^{-(\beta_0 + \beta g_i)}}, \tag{1.8}$$

in this case the vector of outcomes, $\mathbf{y}$, is projected onto the odds scale such that $\mathbf{y} = \frac{f(\mathbf{g})}{1 - f(\mathbf{g})}$. In its simplest case a logistic model will generate $\exp(\beta_1)$ that are

equal to the odds ratio computed using simpler methods previously described. Its main utility is in its ability to incorporate other covariates and nuisance parameters by the addition of terms to the linear predictor.

1.3.5 Bayesian approaches to genetic association testing

Canonical frequentist approaches described in Section 1.3.4 have an important limitation in that their currency, the p-value is unable to capture how confident we are that a SNP is truly associated with a trait (Stephens and Balding, 2009). Conversely a Bayesian posterior provides a strength of evidence for a given hypothesis albeit with additional assumptions and computational burdens. Such posterior probabilities naturally allow the comparison of different hypothesis and lead to a Bayesian interpretation of hypothesis testing, central to this are the concept of Bayes factors.

Bayes Factors

In 1935 Harold Jeffreys described how Bayes theorem could be used to compare competing hypotheses (Jeffreys, 1973). As an illustrative example, consider two hypothesis H_0 and H_1, representing the null and alternative respectively. Given the observation of some data, D, we compute posterior probabilities $\Pr(H_0|D)$ and $\Pr(H_1|D)$. We are interested in the relative support for H_1 compared to H_0 which can be expressed as a quantity known as the Bayes Factor.

$$\text{BF} = \frac{\Pr(H_1|D)}{\Pr(H_0|D)}. \tag{1.9}$$

If we define the prior odds as $\frac{\Pr(H_1)}{\Pr(H_0)}$, then $\frac{\Pr(H_1|D)}{\Pr(H_0|D)}$, the *posterior* odds (PO) can be expressed by

$$\text{PO} = \text{BF} \times \text{prior odds},$$

as set out by Kass and Raftery (1995).

In the context of a case/control setting, let θ be the odds ratio for a given SNP under additive assumptions, and let $\beta = \log(\theta)$, we may assume that $\beta = 0$ and $\beta \sim N(0,W)$ for some specified W under H_0 and H_1 respectively. Thus we construct the Bayes factor as:

$$\text{BF} = \frac{\Pr(D|\beta \sim N(0,W))}{\Pr(D|\beta = 0)}. \tag{1.10}$$

Here, W can be estimated from one of the frequentist approaches previously mentioned.

Wakefields asymptotic Bayes Factors

In three papers (Wakefield, 2007, 2008, 2009) Wakefield introduced an asymptotic Bayes Factor (aBF), for use in association studies, with a simple closed form, requiring as input only the maximum likelihood estimate of $\hat{\beta}$ and its variance V. Such an approach was not only computationally tractable but circumvented the considerable difficulties in obtaining genotype level information, as it could be computed from summary level statistics (e.g. p-values or odds ratios and their standard errors). The aBF relies on a normal prior $N(0,W)$ on β. For a dichotomous trait a value of $W = 0.2$ has been suggested (Giambartolomei et al., 2014) as this approximates to a $\Pr(\theta > 1.4|\theta < 1.4^{-1})$ of 5 %. This prior when combined with the assumption that maximum likelihood estimate of $\hat{\beta}$ is sampled from a normal distribution $N(\beta,V)$ yields:

$$\text{aBF} = \sqrt{\frac{V+W}{V}} \times \exp\left(-\frac{z^2 W}{2(V+W)}\right) \tag{1.11}$$

where z^2 is the Wald statistic $\frac{\hat{\beta}^2}{V}$.

Fine mapping using asymptotic Bayes Factors

The Wellcome Trust Case Control Consortium et al. (2012) show that under the scenario of a single causal variant within a given genomic region containg k SNPs, then the Bayes factor for that region to be causal is the mean of the individual Bayes Factors for all the SNPs in the region,

$$\text{BF}_{\text{reg}} = \frac{1}{k}\sum_{i=1}^{k} \text{BF}_i, \tag{1.12}$$

where BF_i is the Bayes factor associated with the i^{th} SNP being causal, which can be approximated by aBF_i introduced in the previous section. Furthermore they define the single causal variant posterior probability for i^{th} SNP, which substituting for aBF becomes

$$\begin{aligned} \text{sCVPP}_i &\approx \frac{\text{aBF}_i}{k\text{aBF}_{\text{reg}}} \\ &\approx \frac{\text{aBF}_i}{\sum_{j=1}^{k} \text{aBF}_j}. \end{aligned} \tag{1.13}$$

However such a relationship does not consider the case where there are no causal variants within the region. To do this we must alter the definition of BF_{reg}, to include a term for the Bayes factor associated with a model containing no causal variants or $\frac{\Pr(D|H_1)}{\Pr(D|H_0)} = 1$ such that

$$\text{sCVPP}_i \approx \frac{\text{aBF}_i \pi}{\left(\pi \sum_{j=1}^{k} \text{aBF}_j\right) + \pi_0}, \tag{1.14}$$

where π and π_0 are the prior probabilities for a SNP to be causal or not causal respectively. Under the assumption that π is small with respect to k then we approximate $\pi_0 = 1 - k\pi \approx 1$ leading to

$$\text{sCVPP}_i \approx \frac{\text{aBF}_i \pi}{\left(\pi \sum_{j=1}^{k} \text{aBF}_j\right) + 1}. \tag{1.15}$$

Therefore under the strong assumptions, that a given physical region contains a single causal variant, we can define single causal variant posterior probabilities (sCVPP) for each a SNP. By ordering and taking the cumulative sum of these posterior probabilities we can compute credible sets of SNPs that incorporate a given amount of posterior probability (e.g. 95%). In some cases given well powered studies and dense genotyping or imputation this can lead to the identification of single variants upon which the overwhelming majority of the posterior probability is focused, which are then amenable to empirical follow up (Huang et al., 2017).

1.3.6 Approaches to high dimensional data

In previous sections I have focused on inference, however statistical approaches are also concerned with describing and summarising the structure of high dimensional data such as arises from GWAS. Such high dimensional data occurs where the number of observations or predictors, p is much larger than the sample size n, resulting in the so called *curse of dimensionality*; as p increases so does the volume of the space from which our n are sampled, resulting in increased sparsity. If p far exceeds n, as is usual in genomics, then this sparsity is such that robust inference is

challenging. Statistical approaches in this area concern themselves with three main approaches, feature selection, shrinkage (transformation) and extraction (Hastie et al., 2009).

Feature Selection: These are approaches that seek to uncover a subset of the predictors or features that are of relevance to the outcome variable(s). Examples include forward stepwise selection, where predictors are recursively added to a model, and retained only if they improve model performance.

Feature Shrinkage: Shrinkage approaches are concerned with fitting a model with all predictors whilst constraining coefficient estimates. An example of such an approach is ridge regression, which modifies conventional regression approaches (Section 1.3.4), that minimise residual sum of squares, to include a shrinkage parameter, that penalises larger coefficients. Although such an approach improves model performance, its interpretation compared to feature selection methods is challenging as the coefficients of predictors are still non-zero. This has lead to the development of alternative approaches such as the lasso that perform shrinkage, allowing for zero coefficient estimates.

Feature Extraction: This final class of approaches seeks to transform the predictors available in high dimensional datasets in order to derive a much smaller number of features that provide useful summaries of any structure that may be apparent within the data. In most cases, derived features consist of linear combinations of the the original features, therefore reducing the number of dimensions that need to be considered. One of the main approaches is principal component analysis (PCA), which I discuss further in the next section.

1.3.7 Principal component analysis (PCA)

Consider $\mathbf{A}$ a matrix of n observations across p variables where $p >> n$. In such a situation, it is desirable to obtain a representation that summarises $\mathbf{A}$ using a reduced number of variables. PCA is one method that achieves this by concentrating the variance across all p into independent principal components using eigen-decomposition of the covariance matrix, $\mathbf{C} = \mathbf{A}^T\mathbf{A}$. Eigen-decomposition is the process of factorising a square diagonisable matrix, such as $\mathbf{C}$ into an orthogonal set of p eigenvectors, $\mathbf{v}$, and p eigenvalues, λ, such that $\mathbf{Cv} = \lambda\mathbf{v}$. Thus, $\mathbf{C}$ or

indeed **A** when applied to an eigenvector only shrinks or elongates the eigenvector by the magnitude of the corresponding eigenvalue. These eigenvectors are linear combinations of the original variables that are ordered by the amount of variance they capture, the magnitude of which is captured by the corresponding eigenvalue. This allows dimension reduction as we can select a subset of the p eigenvectors, that capture a majority of the variance in **A** to take forward for analysis.

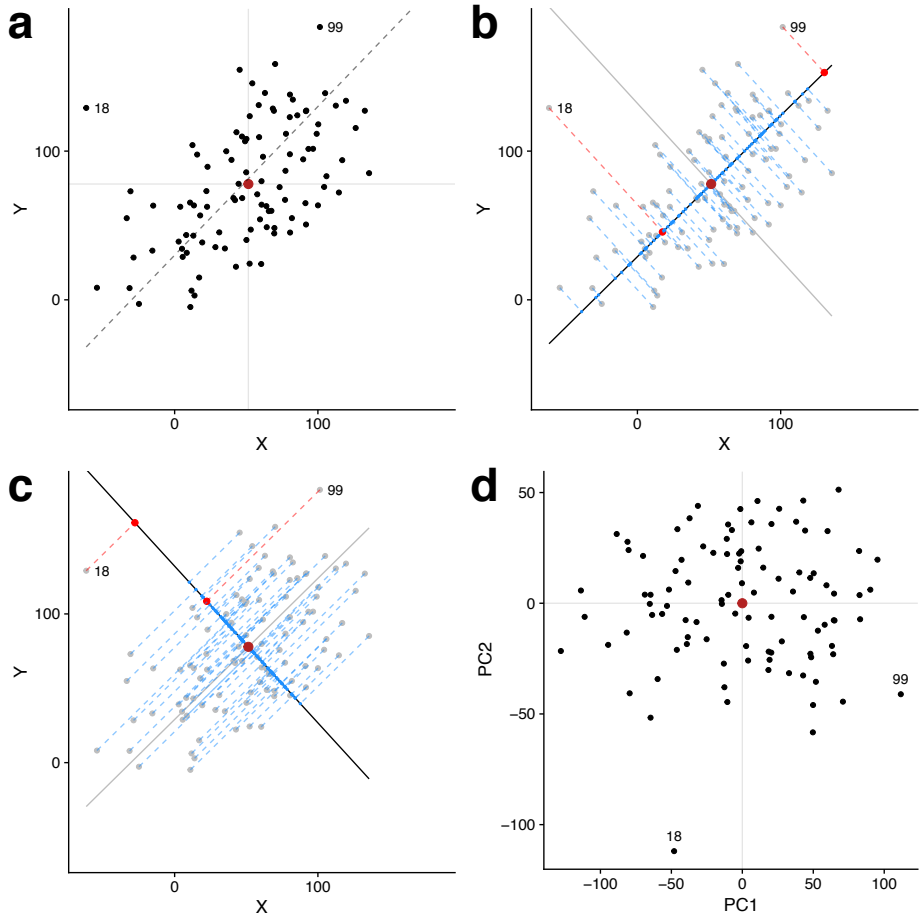


Fig. 1.5 Example of the principal component transformation of a simulated matrix of random variables X and Y ($n = 100$). **a)** Y is a linear function of X (generative function shown as a grey dashed line) with added Gaussian noise, $\varepsilon_x, \varepsilon_y \sim N(0, 0.3)$. The centre of the points, $(\bar{x}, \bar{y})$, is marked in red. Points 99 and 18 are hilighted for illustrative purposes as these are the extreme deviates for X and Y respectively **b)** Transformation shown as dashed lines to the first principal axis, shown as a solid black line. The distance between the centroid and the projection is the principal component score which is maximal for point 99. **c)** Transformation to the second principal axis, here the score for point 18 is maximal. **d)** Biplot showing the principal component scores for both principal axes (blue/red points from **b** and **c**).

Figure 1.5a illustrates this point, showing the simple case of a two-dimensional projection of a set of points onto two possible principal component axes. Here, matrix **v**, becomes the axes, for a new coordinate system or basis. To transform **A** onto this new basis we compute **Av**, illustrated as dashed lines connecting original points to the principal axes (bold lines) in Figure 1.5b (1st principal axis) and Figure 1.5c (2nd principal axis).

For this simple case, we observe that there are two such principal axes or components (as there are two variables), that have two notable relations to each other. Firstly both axis share the same origin ($\bar{x}$,$\bar{y}$), which leads to an alternative viewpoint of PCA as an optimisation problem that seeks to find the set of k dimensions (where $k \leq p$) that minimise the euclidean distance between a set of points. The centroid can be thought of as the zeroth principal axis serving as the origin linking all k dimensions. This constraint enforces the second notable relation, that any k is orthogonal to any other k, such there is no shared variance between any k. As the total variance, the trace of **C**, is constrained this means that each subsequent k captures less variance than the previous. Indeed $\sum \lambda = \mathrm{tr}(\mathbf{C})$, where λ are elements of the vector of eigenvalues obtained from eigen-decomposition of **C** and it is this relation that allows us to compute the ratio of variance explained by each principal axis. The distance of the projection (blue/red points) on the principal axis (bold) from the centroid are the principal component 'scores' for each sample. We can plot these scores in a 'biplot' (Figure 1.5d) to obtain a graphical summary of the sample, here PC1 and PC2 capture approximately 60% and 40% of the total variance respectively.

In this case the PCA is for illustrative purposes only as $n > p$ it has no benefit over a simple scatter plot. Careful thought needs to be given to how input variables are scaled, for example if one variable is measured in metres and another in kilometres then it will appear that the former has a higher variance. As PCA optimises variance between variables this will unintentionally cause the variable measured in metres to have a much greater effect on the basis generated. One method to overcome this is to standardise the variables by mean-centring and dividing through by the variance, thus performing PCA on the correlation rather than variance-covariance matrix (i.e. **C**).

Whilst eigen-decomposition of the variance-covariance matrix, **C** is one method of performing PCA, in practice it is computed through the singular value decomposition of **A** as detailed in Appendix C.1. This approach has multiple benefits in that it does not require the computationally intensive calculation of the variance-covariance matrix, and the loss in precision this can involve.

1.4 Towards causal mechanisms in immune-mediated disease

In the first section I described various high-throughput genomic methods that have been developed to interrogate the context-specific mechanisms by which genes can be regulated with a focus on chromatin organisation. I introduced GWAS and related statistical approaches that have been used to identify regions of the genome harbouring risk SNPs for common complex disease. Here, I expand (Section 1.2.5) on the evidence that integrating GWAS data, both across related traits, and with other sources of high-throughput data is essential to understanding the molecular mechanisms that ultimately modulate complex disease risk. I propose that such knowledge is useful as it could be used to augment or replace disease classifications traditionally based on clinical phenotypes, with more mechanistic ones recognising the blurred lines that often separate related diseases. For the purposes of this thesis I concentrate on a smaller subset of related phenotypes, incorporating autoinflammatory and autoimmune diseases that I unify under the umbrella term immune-mediated disease (IMD).

1.4.1 Epidemiology of immune-mediated disease

A key property of the adaptive immune system is the ability to recognise pathogens from self-antigens. Dysregulation of this process results in damage to healthy tissues, autoimmunity and auto-inflammation. Currently, over 80 diseases have been found to have an underlying immune-mediated pathogenesis, with approximately half presenting as rare diseases, defined as affecting 5 people or fewer in 10,000 (Hayter and Cook, 2012). The collective health burden of more common IMDs such as type 1 diabetes (T1D), rheumatoid arthritis (RA), inflammatory bowel disease (IBD) and multiple sclerosis (MS) is high with approximately 7 - 9% of the European population affected (Cooper et al., 2009). Although environment is a contributing factor in disease susceptibility, the genetic heritability, defined as the proportion of phenotypic variance (Section 1.3.1) attributable to genetic variability, is also important, ranging from 0.39 in primary biliary cirrhosis (PBC) to 0.9 in ankylosing spondylitis (AS) (Gutierrez-Arcelus et al., 2016).

1.4.2 Genetics of of immune-mediated disease

It was in the early 1970's that a role for the human leukocyte antigen (HLA) region was first suggested in ankylosing spondilitis (AS) (Brewerton et al., 1973), and to this day the HLA-B27, class I association with this disease remains the strongest HLA effect of all IMDs. Subsequent work identified a role for the class II HLA-DQβ in modulating risk for T1D (Todd et al., 1987), and in the intervening years associations across a majority of IMDs have been mapped to this region. Such relationships provide significant insight into putative disease processes given that IMDs can be roughly dichotomised as having a predominantly class I or class II disease association (Parkes et al., 2013). Class I HLA associated diseases, which include psoriasis and AS, seldom share associated alleles and are predominantly seronegative, that is they are not associated with specific auto-antibodies. In contrast those characterised by associations in the class II HLA region, which include T1D and systemic lupus erythematosus (SLE), often share risk alleles but associate with specific auto-antibodies (e.g. anti-neutrophilic cytoplasmic autoantibody for SLE) and are thus predominantly seropositive. Furthermore, such seropositive disease are often found to have associations proximal to genes that are the targets of their autoantibodies, for example *INS* in T1D (Bell et al., 1984) and insulin auto-antibodies. In the intervening years with the advent of GWAS, hundreds of regions of the genome have now been associated with one or more IMDs, and detailed analysis supports a central role for both the adaptive and innate immune system in IMD susceptibility, supported by a more limited number of disease-specific loci. Such a wealth of new data has lead to a more systematic analyses of whether associations are shared or distinct between distinct IMDs.

1.4.3 Immune-mediated diseases have both shared and distinct genetic architectures

Evidence for co-morbidity between autoimmune and auto-inflammatory diseases has long been observed. For example, between 4 and 9% of individuals with T1D are also affected by coeliac disease (CEL) (Somers et al., 2006) compared to a general population incidence of 0.8% (Choung et al., 2017). In the era of GWAS it has been possible to examine both genome-wide and locus specific evidence for genetic overlap between immune-mediated diseases. An early study examining whether a genetic overlap could be responsible for the co-morbidity between T1D and CEL found evidence for overlap at 7 out of 28 regions showing

association in either disease (Smyth et al., 2008). Interestingly, two of these overlapping loci, located at 2q12 and 6q25, demonstrated opposing effects between the two diseases. Following up on this, Cotsapas et al. (2011) found evidence for pervasive sharing across seven immune-mediated diseases. These early studies were limited by sample size and the genotyping platforms on which they were constructed, which were conceived to cover, by exploiting linkage disequilibrium in European populations, as much common variation as possible. Without denser genotyping maps it was therefore difficult to elucidate trait-specific fine scale genetic architecture, and thus whether the overlap observed was due to the sharing of causal variants or different variants in genomic proximity.

With this in mind a consortium of 11 immune disease mediated traits, pooled both established and emerging associated regions, mining the developing 1000 Genomes resource (1000 Genomes Project Consortium et al., 2015), to obtain all available variants observed with a minor allele frequency above 1% in European populations local to known IMD associations. This compendium of approximately 186 densely genotyped regions, assaying approximately 200,000 SNPs, formed the backbone for a new bespoke genotyping platform, the ImmunoChip, which could be used to fine map existing associations (Cortes and Brown, 2011) at an unprecedented resolution.

This dense mapping, showed that although there was extensive sharing between IMD risk loci as had been suggested by GWAS, the picture was more complex than initially suggested (Parkes et al., 2013). Whilst associated ImmunoChip regions were shared between diseases, a number exhibited opposing effects, conferring risk in one disease but protection in another. Focusing on T1D, Onengut-Gumuscu et al. (2015) collected summary statistics for an additional 15 IMDs that had been analysed using the ImmunoChip platform into the ImmunoBase resource[3]. Using an enrichment method that controlled for LD, they were able to robustly demonstrate, within T1D-associated ImmunoChip regions, an enrichment pattern that discriminated between antibody positive and negative diseases as previously suggested (Parkes et al., 2013). Fortune et al. (2015) adapted a pre-existing statistical colocalisation method, *coloc* (Giambartolomei et al., 2014), to show robust evidence for the sharing of causal variants at over a third of the associated regions analysed. However the picture of phenotypic and genetic sharing between diseases is complex, for example Fortune et al. (2015), highlighted three regions, associated with T1D but not with other autoimmune diseases, that instead showed

[3]https://www.immunobase.org

overlap with type 2 diabetes. This observation has subsequently been extended to five T1D/T2D regions where evidence suggests a shared causal variant (although in one region opposing effects) by a more recent study (Aylward et al., 2018). Even within IMD, shared genetic architecture is complex, for example psoriatic arthritis and RA show relatively little genetic overlap even though treatment with anti-TNF therapy is effective in both indications (Eyre et al., 2017). Understanding such results biologically is hampered as whilst GWAS are effective at identifying associated variants, elucidating the mechanism by which they act has proved far more challenging.

1.4.4 Integrating functional genomics with GWAS

Having found convincing associations and assessed their overlap between diseases for IMD densely mapped regions, it was clear that a majority of associated variants existed outside of the regions of the genome coding for proteins, a theme common to complex disease (1000 Genomes Project Consortium et al., 2015; Maurano et al., 2012).

In a landmark paper Farh et al. (2015), developed PICS, an attempt to use lead variants retrieved from the GWAS catalogue (MacArthur et al., 2017), in combination with the 1000 genomes reference panel, to fine map association signals in the absence of dense genotyping information. They integrated these results with empirically derived markers for active regulatory chromosome regions across 33 cell types from the Epigenome Roadmap project (Roadmap Epigenomics Consortium et al., 2015). This showed that candidate causal variants were enriched in a trait and cell type specific manner within such regulatory regions. For example, putative causal variants for IMDs were enriched in lymphoid specific regulatory regions, a result that was subsequently confirmed in a study of T1D using credible sets of SNPs (Section 1.3.5) derived from ImmunoChip dense genotype data (Onengut-Gumuscu et al., 2015).

Although this integrative approach could be used to prioritise target cell types and regulatory regions, this was offset against a growing understanding of the complexity underlying cell type specific gene regulation. Importantly this regulatory complexity challenged the established doctrine of the prioritisation of causal candidate genes through the *ad hoc* combination of being physically located with an associated region and having relevant biology (for specific examples see Section 1.2.5).

One systematic approach to link causal variants to their target genes is to elucidate the genetic architecture underlying gene expression, through expression quantitative trait locus (eQTL) studies. Guo et al. (2015) used a statistical method, *coloc* (Giambartolomei et al., 2014), to look for colocalisation between eQTLs identified from a study of three primary immune cell types with GWAS summary results for 10 immune-mediated traits tissues. Interestingly, they found only limited overlap, describing convincing statistical colocalisation in only six genes. A more recent study employing a Mendelian randomisation approach to connect eQTL and GWAS studies also found limited evidence for colocalisation between eQTLs (Zhu et al., 2016) within immune-mediated traits (Huang et al., 2017). Such a lack of colocalisation within IMD, is most likely explained by the fact that eQTL studies are expensive and technically demanding to carry out, limiting not only sample size and thus power, but also the breadth of tissue contexts that can be assayed. More recent studies are beginning to highlight the role of underlying tissue contexts underlying IMD causal variant mechanisms (Alasoo et al., 2018; Soskic et al., 2019).

The overarching conclusion of the integrative studies highlighted is that without an understanding of the tissue context within which causal variants, identified through association studies, act, elucidation of underlying causal mechanisms is challenging.

1.4.5 Towards a mechanistic taxonomy of immune-mediated disease

Whilst, the systematic translation of GWAS results to causal mechanisms has proved challenging, the distinct and shared genetic architectures between different IMDs present an opportunity to organise diseases based on shared molecular phenotypes. Indeed, medicine is to some extent, through the application of diagnostic labels, a systematic categorisation of disease-causing phenotypes. In the past this clustering of individuals sharing clinical features, and pathologies has provided both insight into disease aetiology as well as the development of evidence based treatment options (McCarthy, 2017). As a consequence many diseases are currently classified by their presentation, which in the case of IMDs, is often a proxy for target destruction (e.g. T1D and insulin producing pancreatic β-cells). However in previous sections I have described how, at least at a the level of genetic risk, a more complex picture has emerged, such as in T1D where a central role for adaptive immune system dysregulation interfaces with biological

pathways influencing β-cell function that are themselves shared with Type 2 diabetes. Such pervasive sharing between IMDs (Cotsapas et al., 2011; Fortune et al., 2015; Gutierrez-Arcelus et al., 2016; Li et al., 2015; Parkes et al., 2013) provides scientific support for a new taxonomy of IMDs such as that suggested over a decade ago by McGonagle and McDermott (2006), that recognises that IMDs fall on a spectrum that at its extremities separates auto-inflammatory and autoimmune diseases. However the further development of such a taxonomy depends on first gaining a greater understanding of the underlying causal mechanisms and pathways at work. Such an effort is already well underway, with much progress being in made in the fine-mapping of causal variants (Asimit et al., 2019; Huang et al., 2017; Wallace et al., 2015; Wang et al., 2018), however as set out in the previous section such information needs to be integrated with other sources of genomic information, in order to suggest causal genes and tissues, critical for uncovering disease mechanisms. Uncovering mechanism for individual causal variants is only part of the story, as these then need to be integrated with existing knowledge to identify biological pathways that might be shared across multiple associated regions (Evangelou et al., 2014; Raychaudhuri et al., 2009; Rossin et al., 2011; Võsa et al., 2018). This suggests that approaches at the causal variant scale, as well as more holistic approaches that consider the full spectrum of IMD genetic architectures, are likely to be syntergistic in attempts to develop a therapeutically useful molecular taxonomy of disease aetiology.

1.5 Organisation of the thesis

In this thesis I develop novel methods and apply them to IMD genomic and genetic datasets in order to better understand causal mechanisms and how at the genome level these are shared and distinct between different diseases. In chapters 2 and 3 I develop methods to integrate novel data on three dimensional genome organisation with data from GWAS. These methods allow a data driven prioritisation of putative causal genes and tissue contexts in which they operate. In chapter 4 I develop a PCA dimension reduction technique to in order to summarise the genetic relationships between ten IMDs using GWAS summary statistics as input. Amongst other applications, I investigate whether such summaries have utility in teasing apart the genetic architecture of a clinically heterogeneous juvenile idiopathic arthritis (JIA) cohort, of modest sample size.

1.6 Publications

During this thesis I have contributed to a number of publications which have arisen either directly or indirectly from the work contained in this thesis. Where applicable these are referenced in each relevant chapter, but I list them here for completeness.

- Schofield, E. C., T. Carver, P. Achuthan, P. Freire-Pritchett, M. Spivakov, J. A. Todd, and **O. S. Burren** (2016). CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics 32*(16), 2511–2513.
  Initial concept, co-developed the software and wrote the paper.

- Javierre, B. M.*, **O. S. Burren***, S. P. Wilder*, R. Kreuzhuber*, S. M. Hill*, S. Sewitz,J. Cairns, S. W. Wingett, C. Várnai, M. J. Thiecke, F. Burden, S. Farrow, A. J.Cutler, K. Rehnström, K. Downes, L. Grassi, M. Kostadima, P. Freire-Pritchett,F. Wang, BLUEPRINT Consortium, H. G. Stunnenberg, J. A. Todd, D. R. Zerbino, O. Stegle, W. H. Ouwehand, M. Frontini, C. Wallace, M. Spivakov, and P. Fraser (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell 167*(5),1369–1384.e19.
  Method development and analysis described in Chapters 2 and 3, contributed to writing the paper.

- **Burren, O. S.***, A. Rubio García*, B.-M. Javierre*, D. B. Rainbow*, J. Cairns, N. J.Cooper, J. J. Lambourne, E. Schofield, X. Castro Dopico, R. C. Ferreira, R. Coulson, F. Burden, S. P. Rowlston, K. Downes, S. W. Wingett, M. Frontini, W. H.Ouwehand, P. Fraser, M. Spivakov, J. A. Todd, L. S. Wicker, A. J. Cutler,and C. Wallace (2017). Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biology 18*(1), 165.
  Method development and analysis described in Chapters 2 and 3, co-wrote the paper.

- Inshaw, J. R. J., A. J. Cutler, **O. S. Burren**, M. I. Stefana, and J. A. Todd (2018). Approaches and advances in the genetic causes of autoimmune disease and their implications. *Nature Immunology 19(7)*, 674–684.
  Performed analysis using methods detailed in Chapter 2 to prioritise type 1 diabetes candidate genes.

- Eijsbouts, C. Q., **O. S. Burren**, P. J. Newcombe, and C. Wallace (2019, January). Fine mapping chromatin contacts in capture Hi-C data. *BMC Genomics 20*(1), 77.
  This paper describes an alternative statistical method to CHiCAGO for calling promoter interacting regions resulting in a posterior probability for each interaction. I Investigated whether incorporation of these posterior

probabilities with the method described in chapter 2 could improve candidate gene prioritisation.

- Thaventhiran, J. E. D.*, H. L. Allen*, **O. S. Burren***, J. H. R. Farmery, E. Staples, Z. Zhang, W. Rae, D. Greene, I. Simeoni, J. Maimaris, C. Penkett,J. Stephens, S. V. V. Deevi, A. Sanchis-Juan, N. S. Gleadall, M. J. Thomas, R. B. Sargur, P. Gordins, H. E. Baxendale, M. Brown, P. Tuijnenburg, A. Worth,S. Hanson, R. Linger, M. S. Buckland, P. J. Rayner-Matthews, K. C. Gilmour, C. Samarghitean, S. L. Seneviratne, P. A. Lyons, D. M. Sansom, A. G. Lynch, K. Megy, E. Ellinghaus, D. Ellinghaus, S. F. Jorgensen, T. H. Karlsen, K. E.Stirrups, A. J. Cutler, D. S. Kumararatne, S. Savic, S. O. Burns, T. W. Kuijpers, E. Turro, W. H. Ouwehand, A. J. Thrasher, and K. G. C. Smith (2018). Whole Genome Sequencing of Primary Immunodeficiency reveals a role for common and rare variants in coding and non-coding sequences. *bioRxiv* (Under revision).
  Performed analyses to elucidate novel causes of Primary Immunodeficiency and co-wrote the paper.

* indicates equal contribution.

Chapter 2

Detecting tissue specific enrichment of GWAS signals in PCHi-C data

2.1 Foreword

2.1.1 Chapter Summary

In this chapter I introduce a promoter capture Hi-C dataset (PCHi-C) covering 17 primary haematopoietic cell types (Javierre et al., 2016). I compile a compendium of univariate GWAS summary statistics across 31 traits, proposing a simple LD based method, Poor Man's Imputation (PMI), in order to impute missing data. To assess whether the tissue and trait specific enrichment of GWAS associations reported in the literature (Farh et al., 2015; Onengut-Gumuscu et al., 2015) can be replicated using context specific PCHi-C annotations I develop a novel statistical method, *blockshifter*, that takes into account the correlation inherent in both GWAS and PCHi-C datasets. Through simulation I assess *blockshifter* performance under different assumptions and conditions. I conclude by presenting the results of the application of *blockshifter* to the GWAS compendium.

2.1.2 Attributions

Parts of the work presented in this chapter are included in Javierre et al. (2016) and Burren et al. (2017) and were carried out collaboratively with Fraser, Spivakov, Ouwehand and Diabetes and Inflammation Laboratories as part of the BLUEPRINT/IHEC project (Stunnenberg and Hirst, 2016). Specifically:-

- Dr. Mattia Frontini and Dr. Tony Cutler supplied sorted primary cells.

- Dr. Biola Marie-Javierre, under the supervison of Prof. Peter Fraser carried out all promoter-capture Hi-C experiments.
- Dr. Steven Wingett carried out initial sequence data processing using the HiCUP pipeline (Wingett et al., 2015).
- Dr. Jonathan Cairns and Dr. Mikhail Spivakov supplied PCHi-C chromatin contact maps called using the CHiCAGO pipeline (Cairns et al., 2016).
- Dr. Chris Wallace helped with the conceptualisation of PMI, *blockshifter* and simulation design.

2.1.3 Motivation

In chapter 1 I described how variants associated with common disease, are enriched in non-coding regions of the genome with putative tissue specific gene regulatory function, in a disease relevant manner. One mechanism for the action of such regulatory regions is through the formation of physical chromatin interactions with their target gene promoters. However, the overall relevance of such a mechanism to underlying disease processes is currently unknown. One line of evidence for such a contribution is through the demonstration of the enrichment of disease associated variants in a tissue specific manner, using PCHi-C data. Most published GWAS functional enrichment methods focus on allowing for local correlation structure caused by LD. However, PCHi-C datasets contain significant local structure that must be taken into account. This motivated me to develop a novel statistical method for examining GWAS enrichment in promoter interacting regions (PIRs) that explicitly takes into account this structure, whilst also taking into account LD.

2.1.4 Software availability

PMI and *blockshifter* methods described in this chapter are available from https://github.com/ollyburren/CHIGP and *blockshifter* simulations are available from https://github.com/ollyburren/blockshifter_simulations. Both are available under GNU General Public License v3.0.

2.2 Background

Understanding the tissue context in which disease associated variants function can provide important information for the design of follow up functional experiments.

In Chaper 1, I touched on various analyses that integrate GWAS with tissue specific genome annotations in order to prioritise relevant tissue contexts (Farh et al., 2015; Maurano et al., 2012; Onengut-Gumuscu et al., 2015) which I expand upon here.

2.2.1 Gene set enrichment analysis inspired methods

Early software tools for analysing the enrichment of GWAS association signals in specific functional annotations, built on methodologies such as gene set enrichment analysis (GSEA) (Subramanian et al., 2005), developed to analyse high throughput gene expression datasets. In the context of GWAS, a set of genes with some shared functional annotation is selected, for example, those sharing a common Gene Ontology (GO) term (Ashburner et al., 2000). Variants are then assigned to genes on the basis of proximity, with the most significant association within a fixed genomic window surrounding a gene being taken forward for analysis (Wang et al., 2007). This yields two sets of association p-values; those mapping to genes in the functional annotation to be assayed and the other mapping to a set of background genes, that do not. In the example of Wang et al. (2007) the non-parametric Kolmogorov-Smirnov test is then used to examine whether both sets of association p-values are drawn from the same distribution, to infer enrichment. Recognising the need to adjust for confounders such as gene size and LD, the procedure above is repeated a number of times using a set of association statistics generated from shuffling case/control assignment. This permutation step has drawbacks in that it requires access to the underlying genotype data and is computationally intensive, however it does allow the generation of a well specified null distribution of test statistics which is critical in order to suitably adjust the observed test statistic.

In order to remove the requirement of individual genotyping data, Liu et al. (2010), used the publicly available HapMap reference genotype data set (International HapMap Consortium et al., 2007) to define local LD structure between variants for approximately independent LD blocks. This LD across each block is then used as the covariance matrix, Σ, for a multivariate normal (MVN) distribution, such that $Z \sim \text{MVN}(0, \Sigma)$, which can be used to rapidly generate summary association statistics under the null hypothesis of no association. Such an approach allows the generation, genome-wide, of a background distribution of summary association statistics, that take into account local correlation structure due to LD. This obviates the need for study-specific individual genotyping data, requiring only summary association statistics for the trait of interest, for which enrichment is to be assessed.

Early efforts to characterise enrichment in specific classes of genome annotations, whilst finding plausible enrichment (Maurano et al., 2012), were challenging to confidently interpret. This was mainly due to the reliance on parametric enrichment methodologies that were less robust to local interdependencies within classes of genomic annotations. For example, the base composition of a genomic region may correlate with empirical techniques, underlying an annotation, such as ChIP-seq (Benjamini and Speed, 2012). Such phenomena introduce complicated localised correlation structures within an annotation, that lead to an inflation in enrichment statistics generated if they are not explicitly accounted for.

Another limiting factor was that in order to increase trait coverage by using curated databases of robustly associated trait variants, such as the GWAS catalogue (MacArthur et al., 2017), implicit thresholds were applied, as such catalogues build on findings curated from the literature, rather than the full set of GWAS summary statistics for a given trait. This thresholding not only precluded the incorporation of potentially informative sub-genomewide results (Schork et al., 2013), but also reduced resolution, and depending on the size distribution of annotation classes being examined lead to challenges in non-biased variant assignment.

Another consideration of thresholding is ascertainment bias, or so called winner's curse (Xiao and Boehnke, 2009). Let β be the natural logarithm of the population odds ratio at a given SNP, and $\hat{\beta}$ its sample estimate, it follows that $E(\hat{\beta}) = \beta$, however when a threshold, α', is applied to $\hat{\beta}$ we observe that $E(\hat{\beta} \mid |\hat{\beta}| > \alpha') \neq \beta$, here α' is chosen such that $\alpha = 1 - 2\Phi\left(\frac{|\hat{\beta}|}{\sigma_{\hat{\beta}}}\right)$ is the required significance threshold (where Φ is the integral of the pdf of the normal distribution and $\sigma_{\hat{\beta}}$ is the standard error of $\hat{\beta}$). This occurs in the context of an under powered analysis where the effect size at the declared associated SNP will tend to be more extreme by chance than its true value, conversely, truly associated SNPs, with effect sizes less extreme by chance, will tend not to be declared associated.

Various methods for integrating GWAS with functional annotations that overcome some of these inherent challenges have been suggested. These can roughly be divided into three approaches, which I describe in detail in the following sections.

2.2.2 Matched SNP sets methods

In a matching approach, SNPs in the functional annotation of interest are matched, by various confounding characteristics (e.g. LD and allele frequency) to a control set of SNPs, as exemplified by the GARFIELD software package (Iotchkova

et al., 2019). GARFIELD takes as inputs all univariate p-values from a given GWAS, these are then pruned, using LD information from a relevant reference set of genotypes such that r^2 is less than 0.1 between variants within 1Mb of a trait-associated variant. Functional genomic annotations are assigned to SNPs based on either direct overlap of SNPs in the pruned list or overlap with those in strong LD ($r^2 > 0.8$).

As previously discussed, due to factors including gene length, minor allele frequency and residual LD after pruning, the significance of any enrichments calculated will be inflated. To overcome this an empirical distribution of the enrichment under the null hypothesis, that there is no enrichment, must be computed. To do this GARFIELD repeats the steps described in the previous paragraph on random sets of variants that have been matched to the true set by key metrics including number of LD partners, distance to the closest transcriptional start site (TSS) and minor allele frequency (MAF). Thus, by computing many of these matched null enrichment statistics an empirical p-value can be computed that is adjusted for potential confounders.

2.2.3 Circularised permutation methods

Drawing on insight gained from the ENCODE project (ENCODE Project Consortium et al., 2007), Bickel et al. (2010) rigorously described a 'block subsampling method' which can robustly assess the dependence of one genomic annotation with another. The utility of such an approach, albeit in a simpler form was demonstrated in the analysis of tissue specific gene expression within human regions of differing levels of Neanderthal ancestry (Sankararaman et al., 2014). Trynka et al. (2015) were the first to show that such an approach, exemplified in the GoShifter software, could be used to overcome not only the non-random distribution of genomic annotations with respect to each other, but also the correlation between GWAS signals caused by LD.

In brief the method employed by GoShifter (Trynka et al., 2015), is composed of three steps; Firstly, (Figure 2.1(a)) for a given SNP indexing an association, all variants above a certain LD threshold ($r^2 > 0.8$) are retrieved using a relevant reference genotype set (e.g. 1000 Genomes (1000 Genomes Project Consortium et al., 2015)). Secondly the proportion of annotations which overlap one or more of these variants is computed. Finally the region encompassing all variants identified in initial step is circularised (Figure 2.1(b)). Random rotations of the circularised region are then performed whilst keeping SNP position constant, preserving the

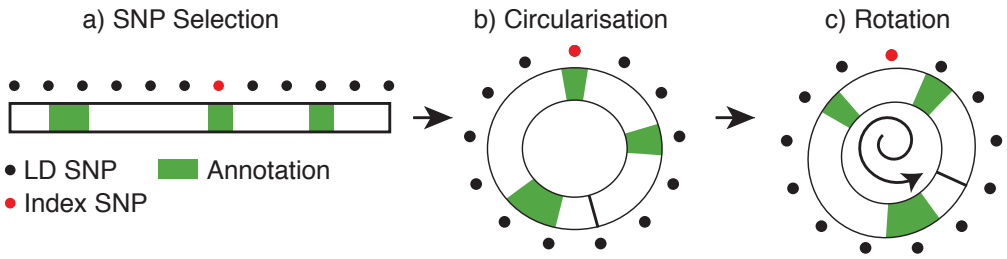


Fig. 2.1 GoShifter method overview; **a)** SNPs in high LD with a SNP indexing the association signal of interest are selected. SNPs overlapping the functional annotation of interest are counted in order to compute an enrichment statistic. **b)** In order to generate a null distribution of enrichment statistics, the LD block defined by variants in **a** is circularised, such that the ends are joined. **c)** This circularised block is then rotated, whilst keeping the positions of SNPs constant, again counting those SNPs that overlap the functional annotation of interest. Step **c** is repeated many times in order to generate the required null distribution of enrichment statistics.

structure of both annotation and variant data (Figure 2.1(c)). For each of these random shifts, the proportion of features overlapping variants is recomputed, which is used to generate a set of empirically derived null statistics. An empirical p-value can then be computed as the proportion of null statistics exceeding the observed overlap proportion.

2.2.4 Statistical modelling methods

The final approach, exemplified by *fgwas* (Pickrell, 2014) and PAINTOR (Kichaev et al., 2014), requires the specification of a joint probability model of variant association and annotation membership based on physical overlap. Whilst the goal of this approach is enhanced fine mapping a bi-product is computation of annotation specific priors or weights that can be interpreted as an enrichment statistic.

Taking the *fgwas* approach as an example the likelihood is specified as follows:

$$\mathscr{L}(y|\theta) = \prod_{k=1}^{M/K} \left(P_k^o (1-\Pi_k) + \Pi_k \sum_{i=1}^{k} \pi_{ik} P_{ik}^1 \right). \tag{2.1}$$

Here K and M are parameters of the approach used to model LD, such that the former represents the number of SNPs in a block and whilst the latter is the total number of SNPs to be considered. P_k^0 is therefore the probability of the observed association statistics (D_k) across a block k under the assumption that there are

no causal variants or $P(D_k|H_0)$. Conversely, P_{ik}^1 is the probability of D_k under the assumption of one causal variant, or $P(D_k|H_1)$. Finally, for the k^{th} block, Π_k and π_{ik} are prior probabilities for the block or the i^{th} SNP to be the causal SNP respectively. Importantly, π_{ik} is modelled as:

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j \in S_k} e^{x_j}}, \tag{2.2}$$

where S_k is the set of SNPs in block k and

$$x_i = \sum_{l=1}^{L} \lambda_l I_{il}, \tag{2.3}$$

where l indexes annotations while L is the number of annotations and I_{il} is set to 1 if the i^{th} SNP overlaps annotation l and zero if not. The relevant parameter for assessment of enrichment of GWAS association signals in an annotation l is λ_l. *fgwas* optimises the likelihood (Equation 2.1), for a given dataset, and in so doing 'learns' values of λ across all input functional annotations, L. Care has to be taken not to overfit the model to the data and *fgwas* employs a penalised regression strategy in order to mitigate this. The likelihood function employed by *fgwas* is hierarchical, and this additional flexibility allows the modelling of an arbitrary number of annotations in an additive manner.

Limitations of existing approaches

Trynka et al. (2015) used simulation to compare GoShifter to the matched SNP set approach and noted that the latter was not well calibrated to control for Type 1 error. They also detected a dependence between choice of matching parameters and observed enrichment. Conversely the circularised permutation approach they employ is based on few assumptions, it does however suffer from the fact that it takes an implicit thresholding approach requiring an input of index variants. As previously discussed (Section 2.2) this will result in missing associations that might be driven by sub genome-wide signals and loss of resolution when applied at the genome-wide scale.

Both statistical modelling and circularised permutation approaches can be used to examine conditional enrichment between annotations, allowing insights to be gained as to how correlation between two sets of distinct annotations (e.g gene TSS and DNaseI hypersensitivity) might be driving enrichment. In theory, matched SNP set approaches could be accommodated to allow for this functionality, in

practice the increase in constraint in selecting matching background sets would make accurate estimation of empirical p-values increasingly challenging.

Both statistical modelling and SNP matching approaches do not make specific provision for localised structure within an annotation. As mentioned it is important to control for intra-annotation dependencies within features of the same annotation type in order to prevent inflation of enrichment test statistics. Indeed, this is a key strength of circularised permutation techniques.

Another consideration is the complexity of the approach. The statistical modelling approaches are the most computationally and operationally complex, especially when multiple annotations are considered. Both GoShifter and GARFIELD are considerably less complex and as a result easier to employ, however this comes at the cost of flexibility in the number of annotations that can be jointly considered. It should be noted that the statistical modelling approaches generate enrichment statistics that are calibrated to be employed as prior probabilities in downstream fine mapping experiments, as such, enrichment is an intermediate rather than an ultimate goal of such approaches.

2.3 PCHi-C maps: description and exploratory data analyses

2.3.1 Tissue coverage

In this thesis I analyse PCHi-C contact maps for a total of 17 primary cell types of the haematopoietic lineage and their empirical derivation as detailed in Javierre et al. (2016). Briefly, each cell type was assessed over at least 3 biological replicates (Table A.1). Promoter interacting *Hin*dIII fragments were called using the CHiCAGO pipeline (Cairns et al., 2016).

2.3.2 Data format description

Contact maps in PeakMatrix format, consist of a baited or sequence captured *Hin*dIII fragment, known as a 'bait' and a list of interacting *Hin*dIII fragments or promoter interacting regions (PIR) within a particular cellular contexts (Figure 2.2). Thus one row represents one interaction, between a bait and PIR. Additional columns map CHiCAGO scores from each of the cell types examined. As the number of bait pair interactions analysed by CHiCAGO is large, only those that

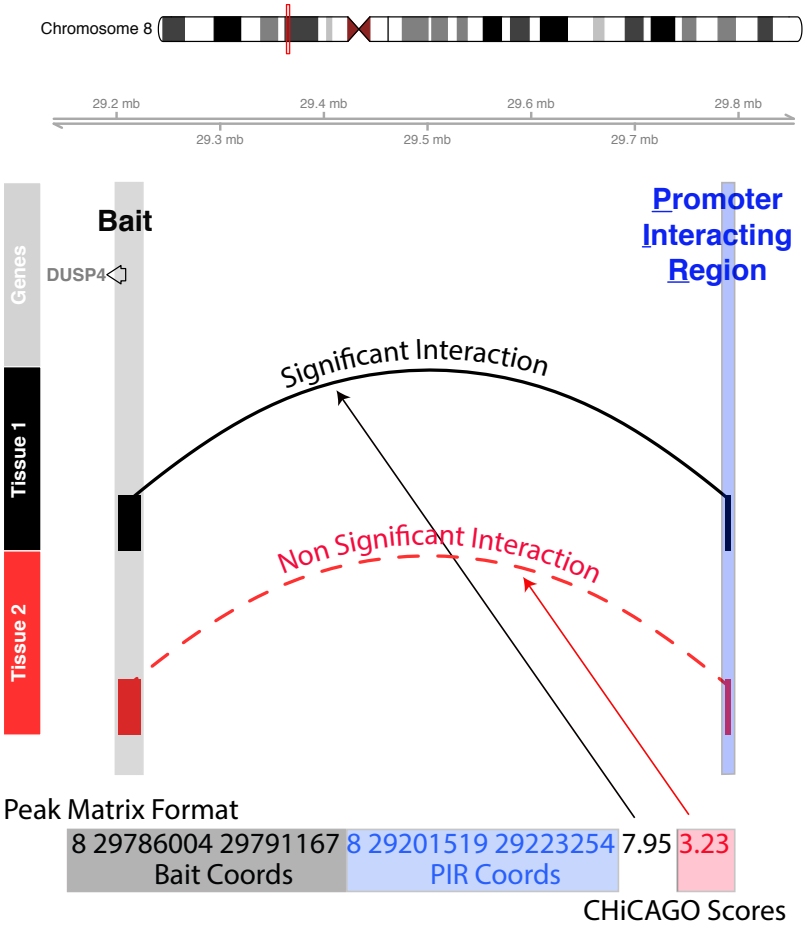


Fig. 2.2 Illustration showing PCHi-C contact map of interactions for *DUSP4*. The promoter capture *Hin*dIII fragment is labelled as 'Bait' and is highlighted in grey. For 'Tissue 1' (black), a significant (CHiCAGO score > 5) interaction with a downstream *Hin*dIII fragment or 'Promoter Interacting Region' (PIR) is shown in blue. In contrast for 'Tissue 2' (red) CHiCAGO does not identify a significant interaction. This information is encoded in one line of Peak Matrix format as illustrated.

show evidence for at least one significant interaction (CHiCAGO score > 5) across the 17 cell types are included.

2.3.3 CHiCAGO score distributions across cell types

I first wanted to understand at a global level, how CHiCAGO scores varied across the different cell types. However, although CHiCAGO scores are quantitative they are challenging to interpret due to their complicated bimodal distribution (Figure 2.3a), with a majority of scores having a score of zero, a second peak around 2 and an extremely large range (0-52). One method, employed in Javierre

et al. (2016) to rescale data with large ranges, is the $\sinh^{-1}$ transform, which reduces the variance and thus range in a non linear fashion. This is desirable so that large CHiCAGO scores do not have undue influence on any inferences made. In general, with this transformation applied the distribution across tissues is very similar (Figure 2.3b), however neutrophils exhibit a different distribution, with many more very low scoring fragments observed, potentially reflecting their unusual segmented nuclear morphology (Javierre et al., 2016).

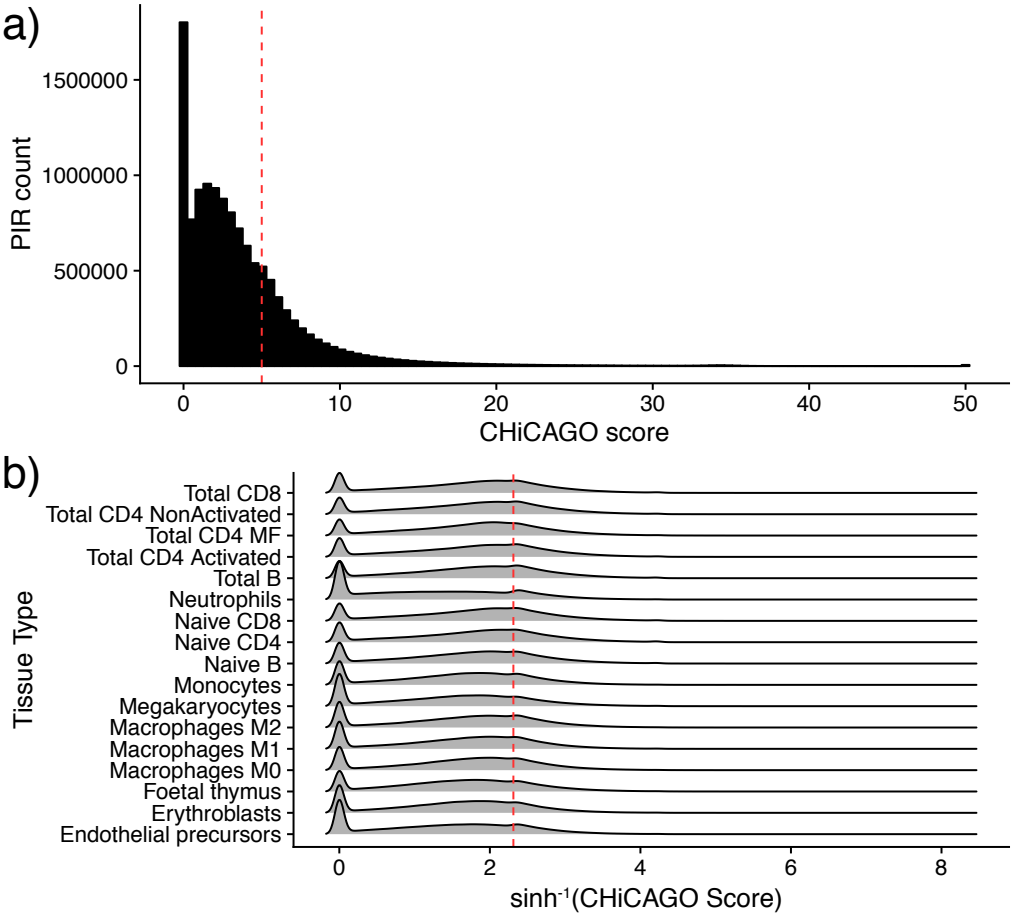


Fig. 2.3 Distribution of CHiCAGO scores. **a)** Distribution of raw CHiCAGO scores, across all cell types. There is an extreme right hand tail distribution, and for clarity scores above 50 are not shown. **b)** A ridge plot showing arcsin(CHiCAGO score) distributions across Cell types. In both panels the red dashed line represents CHiCAGO significance level.

2.3.4 CHiCAGO scores reflect lineage specificity

I used principal component analysis (PCA) on $\sinh^{-1}$ transformed CHiCAGO scores to examine structure within the data. Overall approximately 50% of the

variance was explained in the first two components. Using the PCA derived loadings across all cell types and components I computed the Euclidean distance, using this to perform 'complete' linkage agglomerative hierarchical clustering (See Section 4.4.1). This analysis was able to broadly recapitulate the haematopoetic lineage tree (Figure 2.4) providing support for the biological relevance of this PCHiC dataset. In order to assess the influence of CHiCAGO score thresholding on this result, I carried out a similar analysis setting all CHiCAGO scores less than five to zero, obtaining identical clusters. This indicates that this lineage specific signal is unaffected by employing the recommended thresholding approach and is therefore appropriate for subsequent analyses.

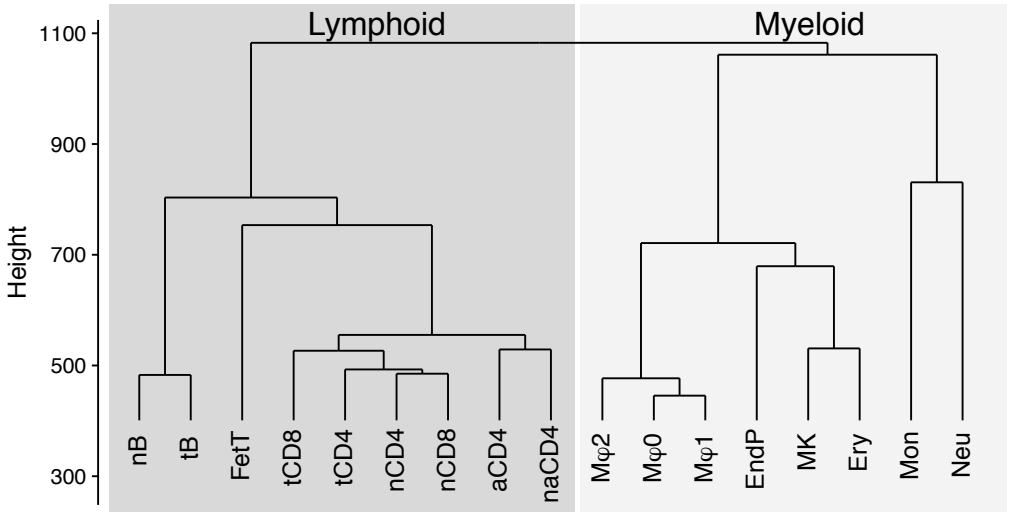


Fig. 2.4 Hierarchical clustering of PCHiC by CHiCAGO score profile that broadly reconstitutes the Haematopoietic tree. **Lymphoid**; nB - naive B cells, tB - total B cells, FetT - Fetal Thymus, aCD4 - activated $CD4^+$ T cells, naCD4 - non-activated $CD4^+$ T cells, tCD4 - total $CD4^+$ T cells, nCD8 - naive $CD8^+$ T cells, nCD4 - naive $CD4^+$ T cells, tCD8 - total $CD8^+$ T cells. **Myeloid**; Mon - Monocytes, Neu - Neutrophils, Mϕ2 - M2 Macrophages, Mϕ1 - M1 Macrophages, Mϕ0 - M0 Macrophages, EndP - Endothelial Precursor cells, MK - Megakaryocytes, Ery - Erythroblasts.

2.3.5 Characterisation of the localised structure within a PCHi-C map

As discussed in the introductory material, for a given functional annotation (e.g. DNase I hypersensitivity) there might be considerable localised structure. I expect this to be a particular issue with PCHi-C datasets, as whilst PIRs might occupy

seemingly disparate physical locations, they are all linked to one or more captured promoter *Hin*dIII fragments. Due to this bait sharing, PIRs are likely to have local structure as a result of both biological and technical phenomena. Local

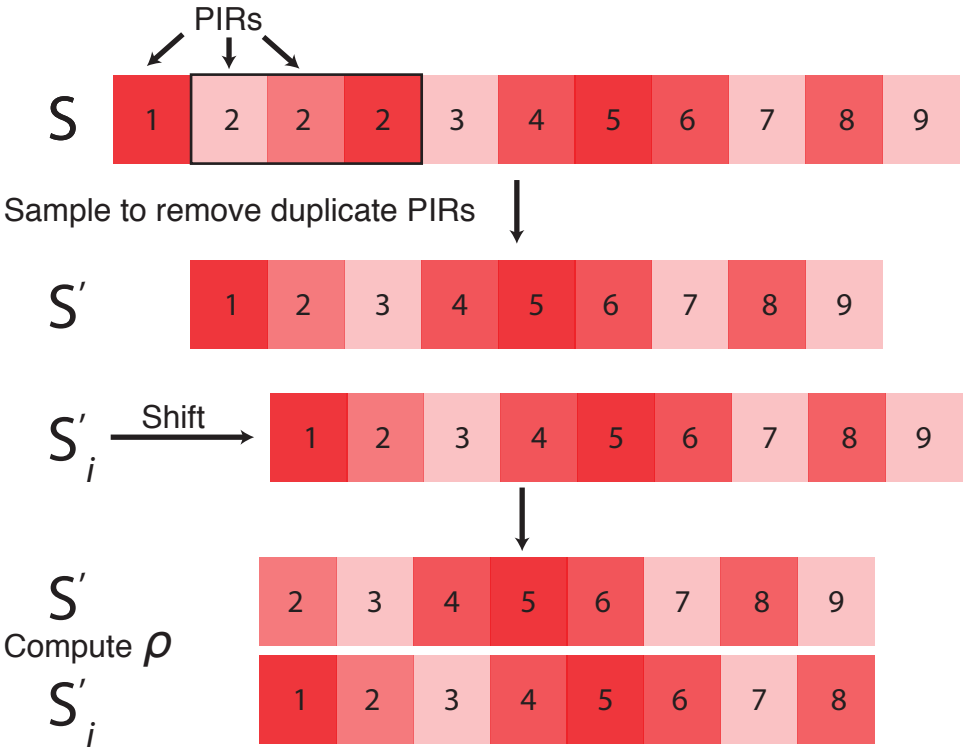


Fig. 2.5 Method for detecting local correlation structure in PCHi-C data. Squares indicate individual PIRs, with pink to red indicating low to high CHiCAGO scores. Duplicate PIRs are removed to obtain $\mathbf{S}'$, which is duplicated and shifted to obtain $\mathbf{S}'_i$. Spearman's ρ is computed between $\mathbf{S}'$ and $\mathbf{S}'_i$, with $\rho \gg 0$ indicating the presence of substantial local structure.

structure will manifest as a spatial correlation between fragment CHiCAGO score and physical separation, which I examined using the following approach. Let $\mathbf{S} = \{s_1, s_2, \ldots, s_m\}$ be a column vector of CHiCAGO scores for m PIRs for a particular cell type, ordered by PIR position (Figure 2.5). Where a PIR has multiple scores (due to the interaction with multiple baits) I randomly sample one score for that PIR in order to prevent duplication to obtain $\mathbf{S}'$. I do this as the method relies on offsetting and duplicate PIRs compromise such an approach. Taking $\mathbf{S}'$ as a reference I shift PIRs by an offset i to obtain $\mathbf{S}'_i = \{s_{1+i}, s_{2+i}, \ldots, s_{m+i}\}$. If $\mathbf{S}'$ is independent from $\mathbf{S}'_i$ and if $i > 0$ we expect $cor(\mathbf{S}', \mathbf{S}'_i) = 0$. As shown in Section 2.3.3, CHiCAGO scores do not follow a normal distribution, therefore I use the non-parametric, Spearman's correlation, taking the mean of the correlation

coefficient ρ as a summary of the correlation structure across cell types. Focussing on chromosome 1 I selected values of i, the fragment offset, that allowed the interrogation of both local and distal correlation structures.

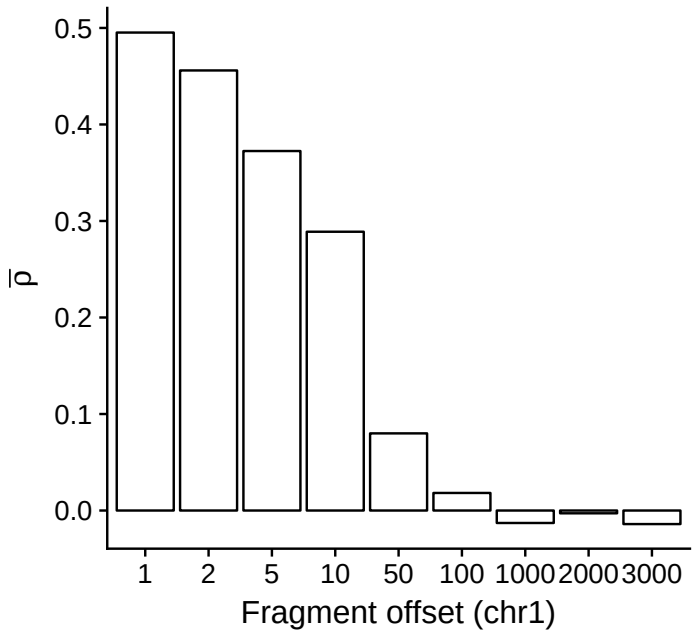


Fig. 2.6 Extent of correlation between CHiCAGO scores and physical location for PCHi-C maps

Across all cell types I observed large amounts correlation at smaller offsets, demonstrating that local correlation structure presents specific challenges when working with PCHi-C datasets (Figure 2.6). This effect decays and is nominal at larger distances reflecting the nature of the underlying biological process being assayed. At extreme offsets above 1000 I see a small negative correlation, that can be explained as stochastic sampling variation in ρ as it approaches zero. This finding of significant local dependence between PIRs underscores the importance of employing an enrichment method specifically designed to mitigate against this in order to generate well calibrated and believable enrichment statistics.

2.4 GWAS compendium: description and data processing

In order to investigate a broad range of phenotypes I downloaded GWAS summary statistics, covering 8 immune-mediated and 23 other traits, from online resources (Table2.1). Genotyping error can create false associations, therefore I filtered all

association statistics to include only robust associations by removing those SNPs which were genome-wide significant ($P < 5 \times 10^{-8}$) but for which no variants within 500Kb and in LD ($r^2 > 0.6$) with the lead SNP existed. Finally I removed any SNP that was genome-wide significant but was not found in the 1000 Genomes Phase III EUR genotype set.

There was a high degree of variation in the coverage of GWAS summary statistics obtained, some studies contained information on approximately 5×10^5 SNPs whereas others were imputed to 1000 Genomes reference genotype set and contained in excess of 7×10^6 SNPs (Figure 2.7). This difference in coverage

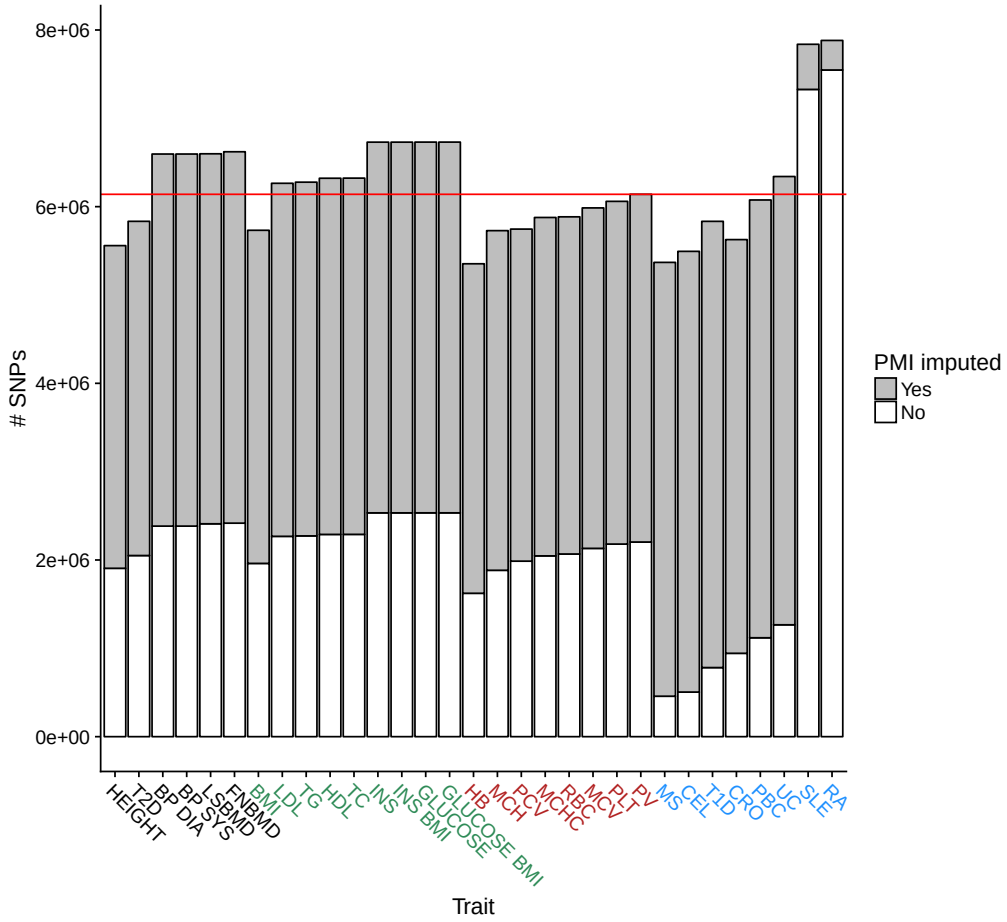


Fig. 2.7 Coverage, pre and post PMI over the Compendium of 31 GWAS traits. Traits are grouped by broad categories, Immune-mediated (blue), Blood (red), Metabolic (green) and Other (black). Grey shading indicates count of PMI imputed SNPs, white counts where data was available from original dataset. The red line is the mean of the total coverage over the entire dataset ($\approx 6 \times 10^6$ SNPs). Trait labels are expanded in Table 2.1.

will negatively impact the ability to annotate association signals to PIRs, which

Trait	Label	Cases	Controls
Multiple sclerosis	MS	9,772	17,376
Celiac disease	CEL	4,533	10,750
Type 1 diabetes	T1D	8,000	8,000
Crohn's disease	CRO	6,333	15,056
Primary billiary birrhosis	PBC	2,764	10,475
Ulcerative colitis	UC	6,687	19,718
Systemic lupus erythrematosus	SLE	4,036	6,959
Rheumatoid arthritis	RA	14,361	43,923
Type 2 diabetes	T2D	12,171	56,860
Haemoglobin	HB	4,627	
Mean corp. haemoglobin	MCH	4,627	
Packed cell volume	PCV	4,627	
Mean corp. haemoglobin conc.	MCHC	4,627	
Red blood cell count	RBC	4,627	
Mean corpuscular volume	MCV	4,627	
Platelet count	PLT	48,666	
Platelet volume	PV	18,600	
Body mass index	BMI	322,200	
Low density lipoprotein	LDL	95,454	
Tryglycerides	TG	96,598	
High density lipoprotein	HDL	99,900	
Total cholesterol	TC	100,184	
Insulin sensitivity	INS	51,750	
Insulin sensitivity BMI adj.	INS BMI	51,750	
Glucose sensitivity	GLUCOSE	58,074	
Glucose sensitivity BMI adj.	GLUCOSE BMI	58,074	
Height	HEIGHT	253,288	
Diastolic blood pressure	BP DIA	69,395	
Systolic blood pressure	BP SYS	69,395	
Lumbar spine bone mineral density	LSBMD	32,961	
Femoral neck bone mineral density	FNBMD	32,961	

Table 2.1 Compendium of 31 GWAS studies analysed. References for each study included are available in Table A.2.

in turn will complicate between study comparisons. Imputation can be used to compute approximate association statistics for missing variants, however, it generally requires access to underlying genotype data, which in this case was unavailable for most traits. Methods exist for imputing summary statistics in the absence of genotyping data such as GCTA (Yang et al., 2011) and IMPG (Pasaniuc et al., 2014). These rely on access to either signed Z scores, odds ratios or β coefficients and their standard errors, in order to estimate direction of effect, which were not available for all studies in the compendium. This motivated me to develop an alternative method to allow the processing of a wide range of traits for which SNP coverage is heterogeneous and only univariate p-values are available.

2.4.1 Defining approximately LD independent blocks

As previously discussed (Section 1.3.1), LD is pervasive in the genome, and thus a single causal variant at a given genetic locus will result in the statistical, but non-causal, association of multiple variants. In order to assess the evidence for variant causality that is comparable across multiple traits it is useful to derive an *a priori* heuristic approach to grouping putative causal variants sharing an association signal. In order to generate approximately independent LD blocks, I used recombination frequency data derived from the CEU sample set of the International HapMap Consortium (International HapMap Consortium et al., 2007) as all GWAS in the compendium were derived from samples of European ancestry. To generate blocks, for each chromosome, I ordered variants for which recombination frequency data was available according to GRCh37 human genome build. I next computed the cumulative sum of recombination frequencies across the ordered variants, for each chromosome, using 1cM boundaries to define groups of variants in linkage disequilibrium. Using these ordered blocks of variants I derived adjacent physical intervals. To do this I selected the variant from the previous block with the largest physical location to define a start, defining the end as the position of the variant in the current block with the maximal physical coordinate. Such a scheme creates adjacent but potentially overlapping physical regions, and to avoid this I added 1 base-pair the start coordinate of each block.

2.4.2 Poor man's imputation pipeline

The poor man's imputation pipeline (PMI) I developed, approximates the p-values for missing SNP summary statistics for a given study using a suitable reference

genotype set. Firstly the genome is split into approximately LD independent LD regions (Section 2.4.1) enabling downstream LD computations to be tractable. For each region I retrieve from a relevant reference genotype set (e.g. 1000 genomes) all SNPs that have MAF > 1% and use these to compute pairwise LD in the 1000 Genomes European population cohort (EUR). The pipeline pairs each SNP with missing p-values to the study SNP with maximum pairwise r^2, r^2_{max}, if that $r^2_{max} > 0.6$, and imputes the missing p-value as that at the paired SNP. SNPs with missing data or without a pair above threshold are discarded, as are SNPs that are included in the study but do not map to the reference genotype set.

2.4.3 Evaluation of PMI performance

Before using PMI in further analyses I wanted to assess the its performance compared to classical imputation, where full genotypes are available. Firstly, I selected all chromosome 1 SNPs from Okada et al. (2014), as a representative sample. This study constitutes a well powered and fully EUR imputed data set. To create a simulated non-imputed data set I pruned these results to contain only SNPs for which p-values were reported in Stahl et al. (2010). I ran PMI on this pruned data set and using bedtools (Quinlan, 2014), aligned the PMI generated p-values with those from Okada et al. (2014). This resulted in a comparison between 235,412 PMI imputed SNPs (Figure2.8).

Generally, there was good agreement between PMI imputed p-values and those derived from classical imputation as reported in Okada et al. (2014) (ρ = 0.94). However, I note that there is considerably more noise in the more modestly associated SNPs for which PMI tends to inflate evidence for association. An explanation for this is that the assignment of p-values based on the PMI method does not take into account the degree of LD. For example a pair of SNPs where $r^2 > 0.99$ will be highly correlated and therefore it is reasonable to expect similar p-values. Conversely, if two SNPs are in more modest LD (e.g. $r^2 = 0.6$) we might expect a decay in association between the two. Such an effect is modelled by classical imputation approaches (used in (Okada et al., 2014)) but not PMI and thus might explain some of the modest inflation in PMI summary statistics detected.

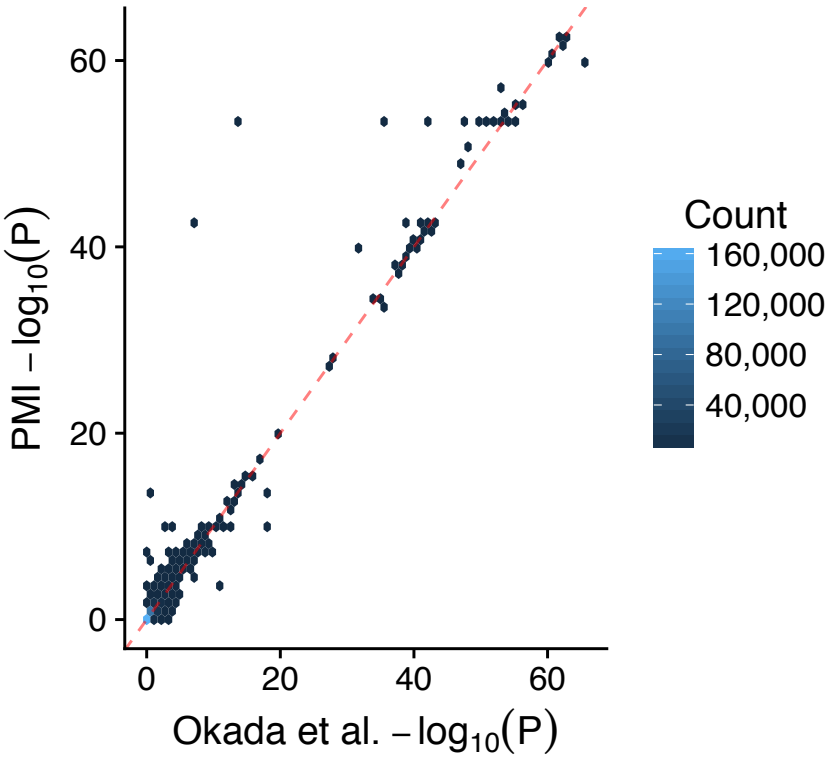


Fig. 2.8 Comparison of p-values imputed by PMI versus those reported in Okada et al. (2014) for Chromosome 1, each hexagonal point represents a bin of SNPs with dark blue and light blue representing low and high counts respectively.

2.4.4 Generation of single causal variant posterior probabilities

As discussed in Chapter 1, univariate GWAS *p*-values are unable to capture how confident we are that a SNP is truly associated with a trait. The asymptotic Bayes factor described in Wakefield (2009) is an attractive alternative as it integrates information about study design and MAF into one metric albeit with the addition of some strong assumptions (Section 1.3.5). The main assumption is that, for a given locus, there exists up to one causal variant. This leads to the question of the best way, genome-wide, to segment the genome, so as to minimise violation of this assumption. One option is to select the set of segments for which LD for a given population is minimised between adjacent segments. Liu et al. (2010) utilised HapMap recombination data in the setting of GWAS pathway analysis to achieve this and in follow up work I had found this to be a suitable segmentation (Burren et al., 2014). Therefore for these reasons and for consistency I used the same LD segmentation employed in the PMI pipeline ($\approx$ 1cM). Another reason for using

this framework is that because the single causal variant posterior probabilities (sCVPP) are derived from a joint model of signals across a given region, LD is naturally taken into account. This mitigates some of the negative effects of PMI, naturally attenuating imputed signals with many LD partners. This occurs because the sum of sCVPP across an LD block is equal to the posterior probability for the region to contain a causal variant (Wellcome Trust Case Control Consortium et al., 2012), which cannot exceed 1. Due to this joint modelling, large signals for variants with many LD partners, exacerbated by PMI, will be equitably shared thus reducing their overall individual impact, naturally reflecting the difficulty of assigning causality to variants within blocks of extended LD.

2.4.5 Prior selection

The computation of aBF requires the selection of two priors. The first is π_i which is expectation of the i^{th} SNP to be causal. I used $\pi_i = 10^{-4}\ \forall\ i$, that is we expect 1 in 10,000 SNPs across the genome to be causal (Giambartolomei et al., 2014). The second prior relates to the value of $\sqrt{W}$, the standard deviation of a normal prior for effect size θ such that $\theta \sim N(0,W)$, which depends on study design considerations. For case/control studies I employ a value of 0.2, this is equivalent to a 95% belief that the true relative risk is in the range of $0.66-1.5$ at any causal variant (Huang et al., 2017). For quantitative trait settings I use 0.15, this is because under the assumption that the outcome variable is normalised such that it's variance is 1, this is equivalent to a variance in trait explained of approximately 0.01 for a SNP with a MAF of 30%. The justification for these priors in discussed further in the supplementary text of Giambartolomei et al. (2014) and Huang et al. (2017).

2.4.6 HLA region

I excluded the HLA region (GRCh37:chr6:25-35Mb) from all downstream analysis due to its extended LD and known strong and complex association with autoimmune disease. Both of these properties would overwhelm subsequent analysis making it difficult to assess whether enrichment was solely driven by the HLA.

2.5 blockshifter development

I sought to develop a method for efficient integration of GWAS data with PCHi-C maps in order to assess enrichment of GWAS associations for each trait across different tissues. Due to the reliance of PCHi-C on *Hin*dIII restriction fragments and the high degree of localised structure in the data, publicly available enrichment tools (as discussed Section 2.2) were unsuitable. Furthermore I wanted to develop a non-thresholded approach based on the causal variant posterior probabilities computed for the reasons outlined in Section 2.4.4.

Considerations

Any tool developed would first and foremost need to be able to account for substantial local correlation structure (Section 2.3.5) that is present in both GWAS and PCHi-C datasets (Figure 2.6). In order to be easily interpretable, any method should compare the enrichment of a set of PCHi-C contacts between two sets of tissues (a test and control set) in a competitive manner. Finally the implementation should allow for the fast computation of enrichment statistics without the need for high performance computing infrastructure.

For global tissue enrichment purposes we are concerned with PIRs and ignore their promoter specific bait linkages i.e. a PIR which interacts with any bait is considered to potentially relate to enhancer presence. This results in considerable duplication as the same PIR can interact with multiple baits. I removed these using the R `duplicate` command prior to subsequent analysis in order to speed up performance.

2.5.1 The *blockshifter* method

I developed a hybrid circularised permutation method (Section 2.2.3) called *blockshifter*, implemented in R, that satisfied the requirements above. This method is competitive in nature and therefore requires user defined sets of test and control tissues for comparison, in addition to pre-computed GWAS study sCVPP$_i$ (Section 2.4.4).

blockshifter uses a statistic, δ, the difference in the mean sCVPP between variants overlapping a set of test and controls PIRs,

$$\delta = \left(\frac{1}{n_1}\sum_{i:i\in \mathrm{PIR_{test}}} \mathrm{sCVPP}_i\right) - \left(\frac{1}{n_2}\sum_{i:i\in \mathrm{PIR_{ctrl}}} \mathrm{sCVPP}_i\right)$$
$$= \left(\frac{1}{n_3}\sum_{i:i\in \mathrm{PIR_{test}}\setminus \mathrm{PIR_{ctrl}}} \mathrm{sCVPP}_i\right) - \left(\frac{1}{n_4}\sum_{i:i\in \mathrm{PIR_{ctrl}}\setminus \mathrm{PIR_{test}}} \mathrm{sCVPP}_i\right), \qquad (2.4)$$

where i indexes variants and $n_1 \ldots n_4$ are the number of variants overlapping a given set. Equation 2.4, shows that PIRs that are found in both test or control sets are uninformative and therefore can be discarded and sufficient statistics to calculate δ in each PIR are $\sum_{i:i\in \mathrm{PIR}}$ and $\mid i : i \in \mathrm{PIR} \mid$ which can be efficiently pre-calculated for a trait and stored.

Whilst the distribution of δ under the null could be estimated empirically, central limit theorem specifies that the normalised sum of a large set independent random variables is approximately normal, even if the random variables themselves are not normally distributed. Such an approach is attractive as estimating the normal distribution under the null, $N(\delta_{\mathrm{null}}, \sigma^2_{\delta_{\mathrm{null}}})$ requires fewer permutations as only, $\mathbb{E}(\delta_{\mathrm{null}})$ and $\sigma^2_{\delta_{\mathrm{null}}}$ parameters require empirical estimation. Local correlation between GWAS signals and between PIRs means that the variance of δ_{null} will be inflated if it naively calculated, as the sum of independent items. It therefore must be estimated through a permutation procedure that preserves local correlation structure

To increase the efficiency of this permutation, runs of one or more PIRs, separated by up to *n Hin*dIII fragments are combined into 'blocks'. These blocks are then assigned into two groups based on test and control set PIR composition. Homogeneous blocks, exclusively containing either test or control PIRs form an 'unmixed' group. Conversely, those with an heterogeneous composition of both test and control PIRs make up a 'mixed' group. Unmixed groups are permuted in a standard fashion by reassigning either test or control labels randomly, taking into account the number of blocks in the observed sets.

The mixed blocks to be permuted require more care and here a modified circularised permutation technique (Bickel et al., 2010; Trynka et al., 2015) is employed. Each mixed block is circularised and the test or control labels of the underlying PIRs rotated (figure 2.9). I store the mean posterior probabilities across each possible permuted block. The number of choices at each block is small, but there are many blocks from which sampling can take place. *blockshifter* then randomly samples from each of these precomputed block permutations n times

so that the proportion of underlying PIRs labels is the same as the observed set and uses this to compute the set of δ_{null}. Finally this distribution of δ_{null} is used to compute an empirical Z-score:

$$Z = \frac{\delta - \overline{\delta_{null}}}{\sqrt{V*}} \tag{2.5}$$

Where V^* is an empirical estimate of the variance of δ_{null}.

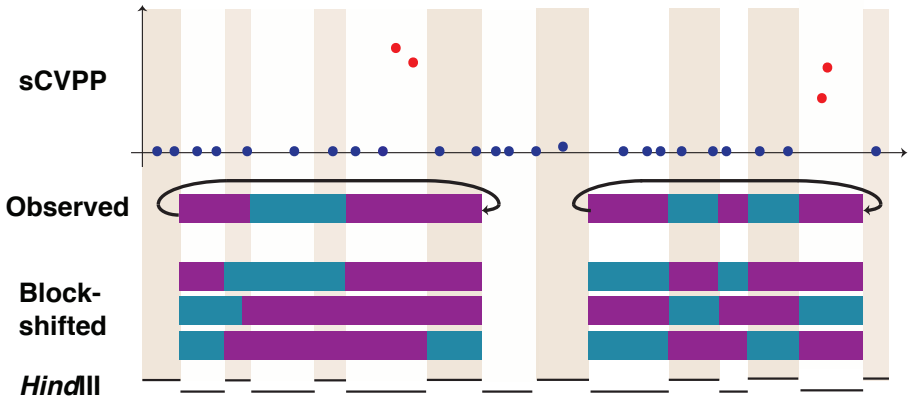


Fig. 2.9 Circularised permutation strategy for two 'mixed' blocks employed by *blockshifter*. GWAS summary statistics are converted to sCVPP (red SNPs have high sCVPPs). Mixed blocks are runs of adjacent *Hin*dIII PIRs found in either 'test' (purple) or 'control' (blue) tissue sets that are separated by n or more *Hin*dIII fragments for which no PIR in either set exists, in this case $n = 2$. sCVPP can be assigned to test or control labels based on PIR overlap and the sum for each set can be stored. The difference in the weighted mean of the sum of posterior probabilities between test and control sets can be used to compute enrichment. However, to control for inflation due to correlation structure between SNPs and between interactions I rotate the labels of *Hin*dIII fragments within the mixed blocks to generate a set of test and control posterior probabilities under the null. Such a strategy is not required for unmixed blocks with a single label. By sampling from these null test statistics across the set of mixed and unmixed blocks (weighted so that I select similar numbers of test and control PIRs to the observed data set) I can rapidly generate an empirical null distribution, genome wide. This can be used to adjust the test statistic to account for inflation due to underlying correlation

2.6 Power and type 1 error rates for *blockshifter*

Before employing any new method it is important to fully characterise its behaviour under simulated conditions. In the case of *blockshifter* I wanted to examine type

1 error rate and it's relationship to a gap parameter indicating the number of missing PIRs allowed within a 'block', which in turn defines the total number of blocks. Intuitively selecting too small a gap size will result in a large number of smaller blocks with a tendency to be of the unmixed block type and the resultant permutation may be insufficient to estimate the variance of δ_{null}. Conversely too large a gap size will result in fewer larger blocks with a tendency to be mixed, where larger scale PIR structure might begin to obscure more local correlation.

As described *blockshifter*, permutes PIR labels whilst maintaining the location of GWAS summary statistics. Any simulation would require a reference test and control PIR dataset that remains constant between simulations. I therefore selected a test (Activated or Non-activated $CD4^+$ T cells) and control (megakaryocyte or erythroblast) set of PIRs with CHiCAGO score > 5, as a reference set for *blockshifter* input.

In contrast I wanted to simulate GWAS summary statistics such that I could control the level of enrichment and use these to understand the performance of *blockshifter* under different scenarios.

2.6.1 Simulation of GWAS

In order to simplify computation I limited the simulation to utilise data for chromosome 1, under the reasonable assumption that this would be representative of genome-wide results. Firstly, I split chromosome 1 into approximately independent LD blocks as described in Section 2.4.2. I used the EUR genotypes to compute a correlation matrix, $\mathbf{\Sigma}$ for variants with a minor allele frequency $> 1\%$, using the `snpStats` R package(Clayton and Leung, 2007). GWAS Z scores can be simulated from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and correlation matrix $\mathbf{\Sigma}$ (Burren et al., 2014; Liu et al., 2010). Each LD block may contain no causal variants ($\text{GWAS}_{\text{null}} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})$) or one ($\text{GWAS}_{alt} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu} \neq 0$).

2.6.2 Enrichment scenarios

For GWAS_{alt}, I picked a single causal variant, i, and calculated the expected non-centrality parameter (NCP) for a 1 degree of freedom χ^2 test of association at this variant and it's neighbours. This framework is natural because, under a single causal variant assumption, the χ^2 at any variant, j, can be expressed as the χ^2 at the causal variant multiplied by the squared correlation, Σ_{ij}, between variants i

and j (Chapman et al., 2003). In each case, I set the NCP at the causal variant to 80 to ensure that each causal variant was genome-wide significant ($P < 5 \times 10^{-8}$). μ, the expected Z score is defined as the square root of this constructed χ^2 vector.

To assess type 1 error rate and power I constructed three artificial scenarios as follows:-

Null enrichment scenario: This scenario, is used to confirm control of type 1 error rate, and causal variants are assigned to PIRs without regard for whether they were identified in test or control tissues.

Moderate enrichment scenario: In order to examine power, the ability of *blockshifter* to reject the null hypothesis in the presence of simulated enrichment, causal variants were preferentially located in test PIRs with a 50% probability.

Strong enrichment scenario: This scenario mimics the preceding scenario with causal variants being located in test PIRs with a 100% probability.

For all scenarios, I randomly chose 50 GWAS_{alt} blocks leaving the remaining 219 GWAS_{null}. Enrichment is determined by the preferential location of simulated causal variants within test PIRs. In all scenarios, each causal variant has a 50% chance of lying within a PIR, to mirror a real GWAS in which we expect only a proportion of causal variants to be regulatory in any given cell type. Note that a PIR from the test set may also be in the control set, thus, as with a real GWAS, not all causal variants will be informative for this test of enrichment.

In order to understand the effect of PMI, for each scenario, I further considered variable levels of genotyping density, corresponding to 'full' genotyping (everything in 1,000 Genomes), HapMap imputation (the subset of SNPs also in Stahl et al. (2010) dataset) or genotyping array (the subset of SNPs on the Illumina 550 k array). Where genotyping density is less than full, I used PMI to fill in Z scores for missing SNPs. I ran *blockshifter*, with 1,000 null permutations, for each scenario and PMI-genotype density condition for 4,000 simulated GWAS, with a *blockshifter* block gap size parameter (the number of contiguous non-PIR *Hin*dIII fragments allowed within one superblock) of between 1 and 20. For all simulations I mirrored the number of cases and controls from the RA dataset (Okada et al., 2014).

For comparison, I also investigated the behaviour of a naive test for enrichment for the null scenario that does not attempt to account for strong local correlation

in both PCHi-C and GWAS datasets. I computed a 2×2 contingency table of variants according to test and control PIR overlap, and whether a variants posterior probability of causality exceeded an arbitrary threshold of 0.01, and used Fisher's exact test to test for enrichment.

2.6.3 Simulation results

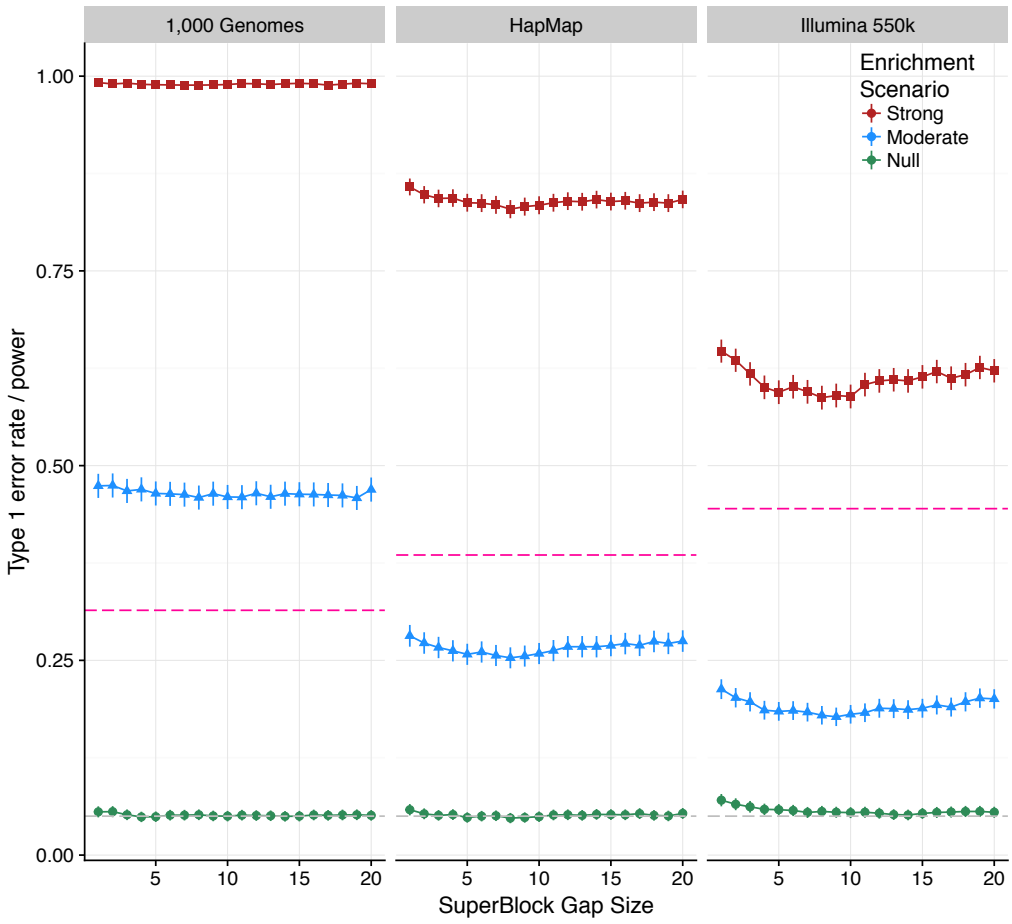


Fig. 2.10 *blockshifter* calibration. Each panel represents a simulated genotyping density: 1,000 genomes (156,082 SNPs); HapMap (44,647 SNPs); Ilumina 550k (10,241 SNPs). Points represent type 1 error rates ($\alpha = 0.05$) for the null scenario (no enrichment of GWAS variants in test specific PIRs) and moderate and strong enrichment scenarios across 4000 simulated GWAS, with differing blockshifter 'block' gap size parameter, n. Error bars represent 95% confidence intervals. Dashed cyan lines represent the type 1 error rate for Fisher's test of enrichment of variants in test and control PIRs under the null scenario.

As expected, the naive application of Fisher's test leads to substantial inflation of type 1 error rate, more so in lower density genotyping scenarios (cyan dashed

lines in Figure 2.10). *blockshifter* maintains type 1 error rate control, although a gap size of 5 or more is required to deal with the extended correlation induced by PMI in lower density genotyping scenarios. As expected *blockshifter* power is attenuated, by the reduction in genotyping density. Furthermore, the effect of applying PMI to impute missing data on *blockshifter* seems to result in a less powerful test with sparser, more PMI imputed values resulting in an attenuated *blockshifter* enrichment statistic (Figure 2.10). Importantly, PMI data do not display evidence for an elevated type 1 error rate.

2.7 Tissue specific enrichment of associated variants with PIRs across 31 traits

I applied *blockshifter* to analyse the compendium of 31 GWAS summary statistics for PCHiC tissue specific enrichment. By design *blockshifter* is a competitive test for enrichment and requires the selection of tissues or tissue groups to compare. A natural dichotomy, in terms of haematopoeitic lineage, exists between lymphoid and myeloid tissues which cluster in two distinct groups as shown in Figure 2.4. Analysis with *blockshifter* showed that variants associated with autoimmune disease traits are enriched at PIRs in lymphoid compared to myeloid cells (Figure 2.11a). A non parametric Wilcoxon rank sum test comparing *blockshifter* Z-scores for autoimmune traits compared to the rest of the compendium showed that this trend was significant (Wilcoxon two-sided $P_{\text{adj.}} = 1.42 \times 10^{-5}$). A further analysis of innate immune cells (monocytes, macrophages and neutophils) compared to megakaryocytes and erythroblasts showed that autoimmune traits were enriched in the former (Wilcoxon two-sided $p_{\text{adj.}} = 7.03 \times 10^{-4}$). There was some evidence that blood traits were enriched in myeloid compared to lymphoid PIRs (Wilcoxon two-sided $P_{\text{adj.}} = 0.005$), however there was no support for this being specific to megakaryocytes or erythroblasts when the two were compared directly which might have been expected. There was no enrichment observed for the metabolic or 'other' categories.

Given the observed enrichment of autoimmune GWAS signals in lymphoid tissues, I performed additional comparisons to see if this could be further localised (Figure 2.11b). The most significant of these was the enrichment of autoimmune GWAS signals in PIRs for $CD4^+$ T cells that had been activated, compared to those that had not (Wilcoxon two-sided $P_{\text{adj.}} = 2.03 \times 10^{-6}$).

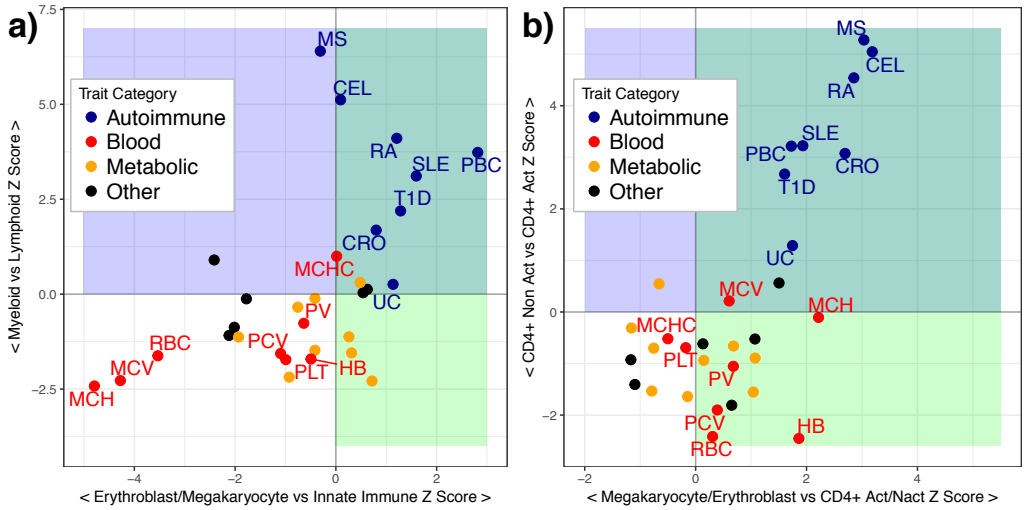


Fig. 2.11 Scatter plot of *blockshifter* enrichment Z scores for 31 GWAS traits. Point colours indicate broad categories of GWAS traits. For clarity, only labels for blood and autoimmune traits are shown (MS - Multiple sclerosis, CEL - Celiac disease, RA - Rheumatoid arthritis, PBC - Primary Billiary Cirrhosis, SLE - Systemic Lupus Erythematosus, T1D - Type 1 Diabetes, CRO - Crohn's Disease, UC - Ulcerative Colitis, PV - Platelet Volume, HB - Haemaglobin, MCH - Mean Corpuscular Haemaglobin, PCV - Packed Cell Volume, MCHC - Mean Corpuscular Haemaglobin Concentration, RBC - Red Blood Cell Count, MCV - Mean Corpuscular Volume, PLT - Platelet Count and PV - Platelet Volume). Each plot represents a set of two competitive comparisons; **a)** horizontal axis Innate Immune Cells (Monocytes,Macrophages and Neutrophils) vs Megakaryocytes and Erythroblasts, vertical axis, Myeloid vs Lymphoid. The blue area represents enrichment for lymphoid PIRs, and the green enrichment for Megakaryocytes and Erythroblasts PIRs. Autoimmune GWAS signals show the strongest enrichment in Lymphoid compared to Myeloid PIRs. **b)** horizontal axis, Megakaryocytes and Erythroblasts vs activated/non-activated $CD4^+$ T cells and vertical axis, activated vs non-activated $CD4^+$ T cells. The blue area represents enrichment for activated $CD4^+$ T cell PIRs, and the green enrichment for combined PIRs across both activated and non-activated $CD4^+$ T cell PIRs.

2.8 Discussion

In this chapter I have described the challenges involved in computing meaningful enrichments for phenotype associated variants with various functional genome annotations. I have shown that by design PCHi-C annotations exhibit significant local structure and this needs to be carefully considered when computing enrichments. Whilst 'off the shelf' solutions exist, none of them explicitly models the restriction fragment nature of PCHi-C datasets and allows the use of non

thresholded GWAS summary statistics, stimulating the development of a novel method.

Whilst extensive catalogues of index variants curated from GWAS exist (MacArthur et al., 2017; Zheng et al., 2017), due to historical privacy concerns (Homer et al., 2008), the full summary statistics encompassing all analysed variants for are a given study are harder to obtain. As for most large scale data analysis projects, a considerable amount of effort was involved in collecting and subsequently conducting quality control on GWAS summary statistics. In the case of immune-mediated disease ,this was ameliorated by the existence of ImmunoBase (https://www.immunobase.org), a publicly available database of curated summary statistics, with a streamlined data access agreement (Onengut-Gumuscu et al., 2015).

In the time since this work was completed other publicly available compendia of summary statistics, with a broader phenotype remit, are gaining momentum. Whilst the Genome-Wide Repository of Associations between SNPs and phenotypes (GRASP) database contains summary data for over 2,000 studies, these are censored such that only variants satisfying $P<0.05$ are included (Eicher et al., 2015). This can be problematic in that coverage will therefore be a function of study size. This when combined with the variable nature of genotyping platforms will result in highly heterogeneous coverage. Of more utility are efforts, exemplified by the GWAS catalogue (MacArthur et al., 2017), to compile complete summary statistics over a broad range of traits under light touch 'click-through' data access agreements. These efforts, if supported, have the ability to significantly facilitate integrative types of analysis such as that described here.

Once summary statistics are available, due to nature of the underlying genotyping platforms utilised, coverage is heterogeneous (Figure 2.7). To a certain extent some of this is historical, as extensive infrastructure (Das et al., 2016; McCarthy et al., 2016) has been developed in order to simplify the task of imputing genotype data to high density reference haplotype datasets. Nevertheless, datasets currently available do suffer from this heterogeneity. It was therefore necessary to develop PMI in order to overcome this.

Through simulation I was able to show that *blockshifter* was increasingly conservative and that power was attenuated as genotyping density decreased, increasing the number of PMI imputed SNPs incorporated. One concern is that two diseases (multiple sclerosis and coeliac disease) with the lowest genotyping density (Figure 2.11) exhibit the most enrichment. One explanation for this is that the simulation of GWAS summary statistics is non-trivial under the alternative and

simulated causal variants were assumed to have GWAS *p*-values that were above genome-wide significance. Further work could be undertaken to use more up to date GWAS simulation frameworks, such as simGWAS (Fortune and Wallace, 2018) to further investigate the performance of *blockshifter* under sub-genomewide significant causal variant conditions in order to understand whether this might cause significant inflation. It should be noted, however, that rheumatoid arthritis and SLE, two large studies imputed to 1000 Genome reference still exhibited enrichment supporting the overall finding for enrichment across the ensemble of immune-mediated/autoimmune diseases studied.

The primary motivation behind the development of *blockshifter* was to ascertain whether I could integrate a large PCHi-C dataset with GWAS summary statistics to obtain biologically relevant enrichments. Given that the chromatin interactions assayed by PCHi-C are enriched for enhancer elements (Javierre et al., 2016), failure to replicate, robustly, a global enrichment for GWAS signals in a relevant tissue specific manner would discourage the further development of methods to integrate GWAS and PCHi-C datasets. Generally given the haematopoieitic focus of the PCHi-C dataset, I detected broadly expected enrichment patterns with autoimmue disease enrichment in lymphoid rather than myeloid tissue (Onengut-Gumuscu et al., 2015). In contrast, blood traits such as mean corpuscular haemoglobin showed enrichment in erythroblast/megakaryocyte compared to innate immune cells such as macrophages and monocytes. It was interesting that enrichment was not detected for platelet volume or platelet count, given the precursor role of megakaryocytes in the genesis of platelets. One explanation is the growing evidence suggesting a role for platelets in both adaptive and innate immune function (Morrell et al., 2014), as such there might be unappreciated overlap in cellular programs underlying megakaryocyte and adaptive and innate immune tissues, resulting in no overall enrichment in this competitive setting.

Chapter 3

Integrating GWAS and PCHi-C data to prioritise causal genes and tissues

3.1 Foreword

3.1.1 Chapter Summary

In this chapter I develop a method, capture Hi-C omnibus gene score (COGS), to integrate genetic and chromatin conformational data in order to prioritise causal genes and tissues for GWAS traits. I employ COGS to prioritise candidate causal genes across the 17 PCHi-C and 31 GWAS datasets described in Chapter 2. To characterise its utility I compare COGS prioritisation results using PCHi-C, conventional Hi-C and proximity-based input methods. To identify putative tissue specific mechanisms I develop a heuristic framework, hierarchical COGS which I then apply to the compendium. Focusing on activated and non-activated $CD4^+$ T cells, I analyse ImmunoChip data across four autoimmune diseases and compare the effect on COGS performance of fine mapping approaches assuming either single or multiple causal variants in approximately LD-independent genomic regions. Finally, I integrate orthogonal functional annotations to refine prioritised genes and describe a biological validation in *IL2RA*, by a collaborator, in support of the approach.

3.1.2 Attributions

Parts of the work presented in this chapter are included in Javierre et al. (2016), Burren et al. (2017) and Inshaw et al. (2018) and were carried out collaboratively with Fraser, Spivakov, Ouwehand and Diabetes and Inflammation Laboratories as part of the BLUEPRINT/IHEC project (Stunnenberg and Hirst, 2016). Specifically:-

- Ms Ellen Schofield helped me to co-develop https://www.chicp.org (Schofield et al., 2016), a resource to visualise PCHi-C interaction maps in the context of GWAS association statistics. Whilst this work is outside of the scope of this thesis, it was a useful tool in checking the validity of many of the analyses described in this chapter.
- Dr. Csilla Varnai, Mr. Michiel Thiecke and Dr. Mikhail Spivakov provided TAD annotations used in COGS input method comparison (Section 3.6.1).
- Dr. Roman Kreuzhuber, under the supervision of Dr. Oliver Stegle carried out eQTL analysis (Section 3.6.4).
- Dr. Chris Wallace helped with the conceptualisation of COGS (Sections 3.4 and 3.7.1) and supplied GUESSFM results (Section 3.8.4).
- Dr. Antony Cutler, Mr. Arcadio Rubio-Garcia and Dr. Chris Wallace provided analysed ChIP-Seq, RNA-Seq data and eRNA annotations for activated and non-activated $CD4^+$ T cells (Section 3.8.5).
- Mr. Daniel Rainbow and Dr. Chris Wallace provided allele specific expression data and analysis for *IL2RA* (Section 3.8.6).

3.1.3 Motivation

In the previous chapter I demonstrated that disease associated variants, identified through GWAS, were enriched in PCHi-C identified PIRs in a tissue specific manner, thus providing further support for their role in mediating disease risk. However such a global analysis has limited value in suggesting specific causal mechanisms for functional follow up.

By definition, each tissue specific PIR, identified by PCHi-C, is paired with a gene promoter. This motivates further investigation as to whether they can be used to link variants which causally affect disease risk to their target genes in specific tissues. Such a data-driven approach to causal gene prioritisation might improve

the efficacy of downstream functional characterisation, by providing a ranking of not only causal genes but also relevant tissue contexts.

3.1.4 Software availability

An implementation of the COGS method as described in this chapter is available from https://github.com/ollyburren/CHIGP under GNU General Public License v3.0. A more portable and recent R package version, including a vignette, is available from https://github.com/ollyburren/rCOGS under an MIT licence. Collaboratively, I co-developed a browser based tool with Ellen Schofield to visualise PCHi-C results in the context of GWAS summary statistics (Schofield et al., 2016) which is publicly accessible at https://www.chicp.org (full source code is made available from https://github.com/ollyburren/django-chicp).

3.2 Background

In the introductory material I described how GWAS has been successful in identifying complex disease associated variants (MacArthur et al., 2017). However, integration with functional annotations has demonstrated that many of these signals map to tissue specific regulatory sequence elements (Dendrou et al., 2016; Farh et al., 2015; Maurano et al., 2012; Onengut-Gumuscu et al., 2015) affecting the regulation of gene expression rather than protein coding sequence. Frequently, through 'chromatin looping', these regulatory elements interact with target gene promoters over large physical distances, often 'skipping' upstream or downstream intervening gene promoters, presenting a significant challenge for variant interpretation. Whilst targeted functional characterisation of individual GWAS loci is essential (Claussnitzer et al., 2015; Davison et al., 2012; Smemo et al., 2014), systematic, data-driven approaches have utility in narrowing the hypothesis space by prioritising putative causal variants, genes and importantly the tissue contexts through which they act.

3.2.1 LD and Proximity approaches

In the early days of GWAS, where emphasis was placed on locus discovery and replication, resources were directed to enlarging sample collections and developing complimentary methods for efficient and robust data analysis (Clayton and Leung, 2007; Purcell et al., 2007). This, combined with a paucity of information about the

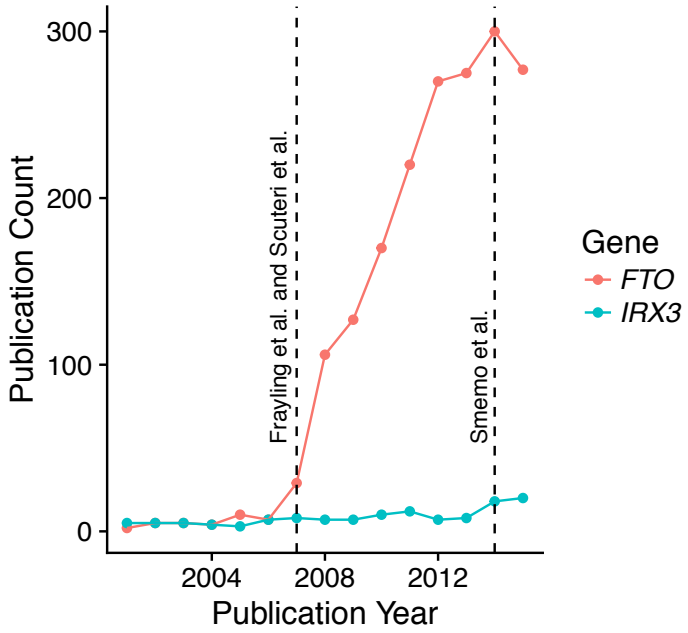


Fig. 3.1 Publication count each year, between 2001 and 2015, for *FTO* and *IRX3*. Dashed vertical lines indicate pioneer studies suggesting a putative causal role for *FTO* (Frayling et al., 2007; Scuteri et al., 2007) and *IRX3* (Smemo et al., 2014) in obesity.

function of non-coding variants in relevant tissue contexts, led to the application of expert-based heuristic methods in order to suggest causal candidate genes. A popular framework has been to use LD or recombination data to compute a genomic locus within which the causal variant or variants reside, intersecting these intervals with protein coding gene annotations in order to create lists of candidate genes. Often, such long lists, are then refined by domain experts in order to prioritise putative causal genes for functional follow up experiments. As discussed (Section 3.2), regulatory variation can act over large distances, making proximity an unreliable indicator for causal gene candidacy. This, combined with a reliance on literature and expert knowledge, for which knowledge of genes and their biological contexts is often heterogeneous, can obscure the candidacy of poorly characterised genes. This can in turn, feedback into the literature, as the mention of a putative causal disease gene in an influential publication can have a stimulatory effect on it's future study (Figure 3.1). This self reinforcement of the association of a particular gene with a disease can draw focus away from other candidate genes, inadvertently retarding disease research progress. Such difficulties in drawing reliable biological insights from GWAS have challenged the community (Bahcall, 2012) to shift focus away from locus discovery towards

developing methods that integrate both genetic and genomic information in order to suggest causal mechanisms in a more data-driven, hypothesis-free fashion.

3.2.2 Population genetics approaches

One of the main data-driven approaches for linking causal variants to target genes and relevant tissue contexts is to harness population genetic approaches, that examine how molecular genomic quantitative traits correlate with genetic variation. Such studies were pioneered in humans, by the analysis of the modulation of gene expression by genetic variation (Stranger et al., 2007). Typically such expression quantitative trait loci (eQTL) analyses, begin by the collection of gene expression measurements for hundreds of genotyped individuals for the cell type of interest, either using microarray, or more recently, RNA-Seq analysis techniques. This catalogue of n genotypes and p expression levels can be combined in a series of univariate linear regressions to identify those variants that are significantly associated with target gene expression. In the naive setting one could consider a genome-wide analysis where each available variant is regressed against each gene ($n \times p$ tests). In practice such a strategy is not employed, because of the significant multiple testing burden incurred. Since target gene proximity is related to eQTL frequency (Brem et al., 2002) analysis is instead restricted to detecting *local* eQTLs (Albert and Kruglyak, 2015), such that analysed variants are limited to those overlapping a fixed physical window centred on the gene of interest, resulting in many fewer tests.

In humans the detection of such local eQTL's in a multitude of tissue contexts has been successful, with a recent study, GTEx, robustly detecting one or more local eQTLs for 86% of protein-coding genes across 44 post-mortem tissue types (GTEx Consortium, 2017). Identifying overlapping eQTL and GWAS signals provides a powerful and intuitive mechanism for prioritising genes and tissue contexts for follow up of GWAS results. In the face of ever broadening tissue catalogues of eQTLs, interpreting such overlaps has proved challenging. For example, GTEx found that 93% of all common variants are nominally ($P < 0.05$) associated with expression of at least one gene in one or more of the tissues that they studied (GTEx Consortium, 2017). Indeed the application of more robust statistical methods (Giambartolomei et al., 2014; Zhu et al., 2016) designed to overcome the effect LD plays in such overlaps, have found relatively limited robust evidence for overlap (GTEx Consortium, 2017; Guo et al., 2015; Zhu et al., 2016).

The powerful eQTL paradigm described is extensible to any high throughput 'omic' technology that can be used to robustly generate quantitative molecular phenotypes. Indeed it is now routinely used to assess a multitude of other functional annotations such as chromatin accessibility (Alasoo et al., 2018; Tehranchi et al., 2016). Whilst genotyping need only be carried out once per individual, molecular phenotyping needs to be repeated for each trait and tissue permutation of interest. In practice, due to understandable economic considerations this limits not only cohort size and thus power to detect variants with more subtle effects, but also those functioning in specific tissue contexts relevant to disease aetiology (Rockman, 2012).

3.2.3 High throughput molecular genomic approaches

Initial human genome-wide catalogues of chromatin modification elucidated that regions associated with complex disease were enriched for modifications marking regions of regulatory activity using correlation between disparate elements in order to infer target effector genes (Farh et al., 2015; Maurano et al., 2012; Thurman et al., 2012). As discussed in detail in Chapter 1, advances in high throughput chromatin conformational capture (Dekker et al., 2002) along with subsequent sequence capture extensions, have allowed the consideration of specific disease loci in three-dimensions and the physical linkage of putative causal variants with their target gene promoters (Dryden et al., 2014; Martin et al., 2015). Further development of high resolution PCHi-C has facilitated the genome-wide analysis of physical linkages between GWAS association signals and their target genes (Mifsud et al., 2015). However, none of these previous studies have combined the statistical fine mapping methods described in Chapter 1 with PCHi-C data across disease relevant cell types to infer causal linkages systematically.

3.3 Promoter-capture platform coverage

Throughout this chapter as the source of PCHi-C promoter interacting regions (PIRs), I use data from Javierre et al. (2016) as described in Chapter 2. The PCHi-C platform employed uses the same capture library as described in Mifsud et al. (2015), originally designed to capture the promoters of 89% of Ensembl protein-coding genes, noncoding RNAs, antisense RNAs, small nuclear RNAs (snRNAs), microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs) available at that time. Complete coverage is not possible due to the technical challenges in

designing probes to capture all promoter containing *Hin*dIII fragments, such as those in regions of the genome where there is a high degree of repetitive sequence.

3.3.1 Capture platform reannotation

Any approach linking putative causal variants to genes via PCHi-C and GWAS integration will benefit from a thorough and up-to-date integration of gene promoter locations with *Hin*dIII fragment capture design.

I therefore re-annotated the underlying promoter sequence capture platform (Mifsud et al., 2015) using Ensembl version 75 (Yates et al., 2016) gene annotations. I defined a promoter as the physical location immediately 5' to a gene transcriptional start site (TSS). As genes may consist of multiple transcripts this results in a many-to-many mapping of promoters to fragments. Assignment of genes to fragments was conducted at the gene level and multiple TSS for the same gene in the same capture fragment were collapsed.

3.3.2 Distribution of captured transcriptional start sites

In summary 22,076 captured *Hin*dIII fragments (baits), containing 31,253 non redundant promoters were annotated, covering 18,202 protein-coding and 10,928 non-protein coding genes (Table 3.1). This resulted in a 90% coverage of protein coding space of at least one promoter per gene. There was no evidence to support a systematic failure to capture genes in specific genomic loci such as telomeric regions (Figure 3.2).

The underlying resolution of PCHi-C is dependent on the distribution of *Hin*dIII cut sites within the human genome, resulting in a mean and median length of approximately 9Kb and 7Kb for captured fragments respectively (Figure 3.3a). As expected I found a relationship between fragment size and the number of promoters captured, with larger fragments capturing more promoters (Figure 3.3b).

As previously mentioned, a single bait might also contain multiple promoters; these can include promoters for alternative transcripts for the same gene or more problematically, promoters for multiple genes. Setting aside non-protein coding genes, I found that approximately 18% of captured fragments fell into the latter category of what I term, *promiscuous* baits. This rises to 33% on the incorporation of non-coding classes of genes.

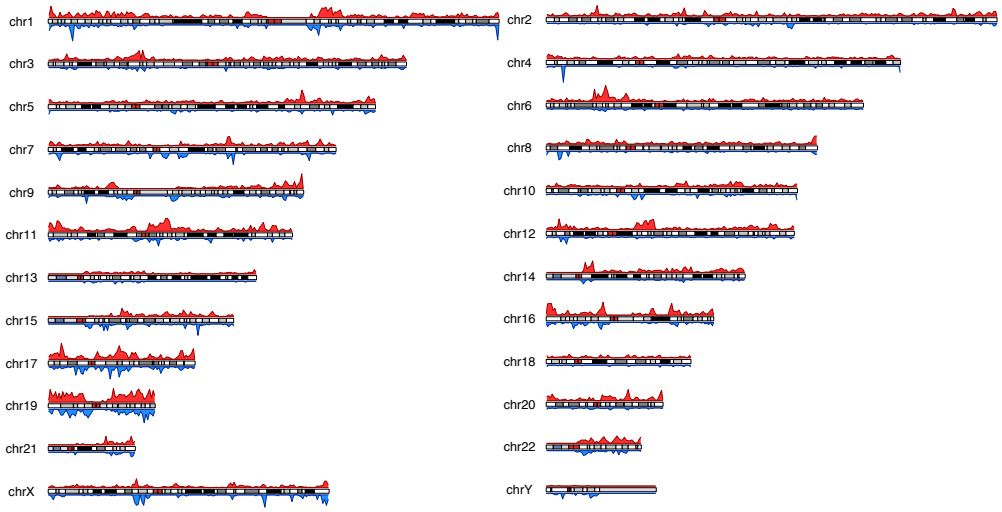


Fig. 3.2 A karyotype coverage plot showing captured (red) and missing (blue) protein coding gene promoters for the PCHi-C platform employed.

Gene Biotype	Overall	PCHi-C Bait	Coverage (%)
protein coding	20,314	18,202	90
snoRNA	1,457	1,286	88
snRNA	1,916	1,571	82
antisense	5,273	3,782	72
miRNA	3,049	1,532	50
processed transcript	514	249	48
polymorphic pseudogene	45	13	29
sense overlapping	202	58	29
sense intronic	741	145	20
3′ overlapping ncRNA	21	4	19
lincRNA	7,109	1,008	14
IG C pseudogene	9	1	11
misc RNA	2,033	230	11
rRNA	526	42	8
pseudogene	13,920	994	7
IG V gene	138	5	4
IG V pseudogene	187	8	4

Table 3.1 PCHi-C sequence capture by Biotype using Ensembl v75 gene annotation. 'Overall' - Total number of features annotated in reference genome. 'PCHi-C Bait' - Total number of features captured in PCHi-C bait design.

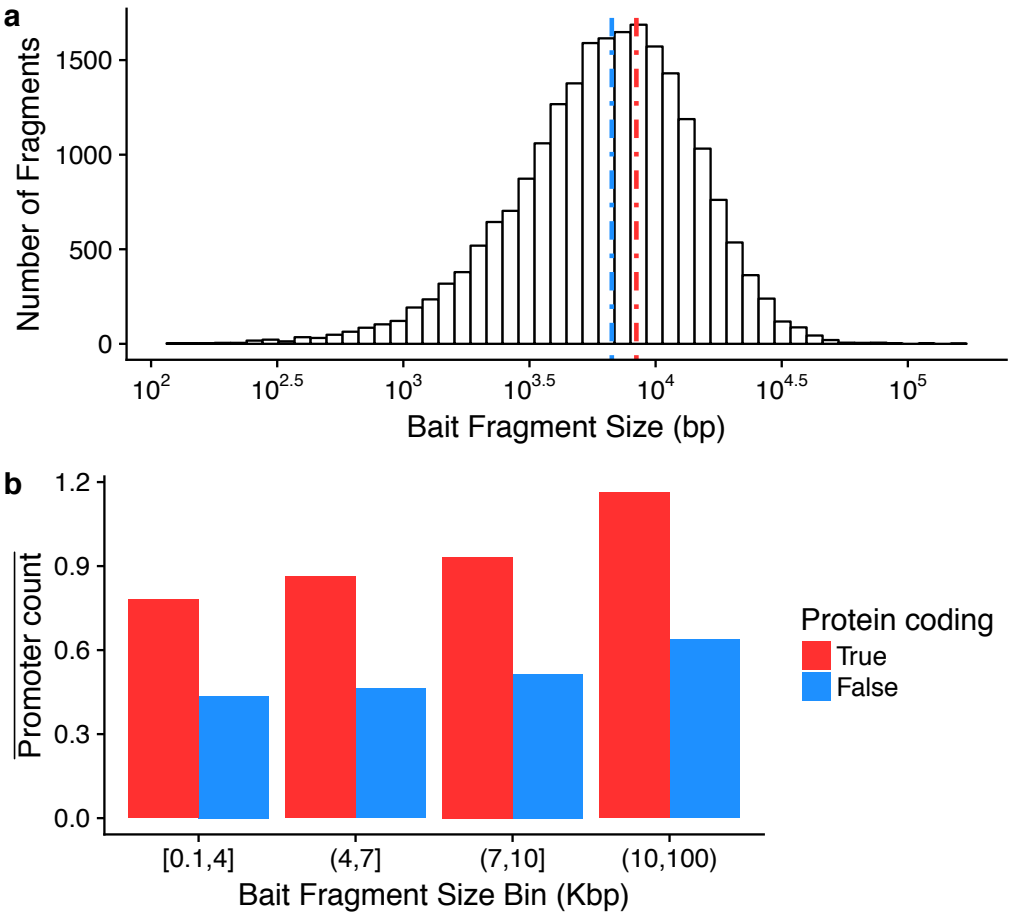


Fig. 3.3 Captured *Hin*dIII fragment sizes and promoter overlap distributions; **a)** distribution of baited fragment sizes with blue and red broken lines indicating median and mean fragment lengths respectively, **b)** mean promoter counts for quantile bins of bait fragments stratified by promotor biotype.

3.4 A method to integrate GWAS summary statistics with PCHi-C maps

3.4.1 Method overview

In order to integrate GWAS data with data linking genomic regions to gene promoters I developed a gene score metric, Capture Hi-C Omnibus Gene Score (COGS). In overview this metric uses trait specific sCVPP (Section 1.3.5), in conjunction with PCHi-C maps in order to assess the evidence that a gene is causal in specific gene and tissue contexts.

To compute an overall score for a gene I consider evidence for association of the GWAS trait to variants across three broad sets of genomic features: coding,

'virtual' promoter, and PIRs (derived from, for example, PCHi-C), within a given approximately LD independent region (Section 2.4.1).

I quantify 'evidence for association' using the sCVPP. In a gene/tissue prioritisation setting and in the absence of individual level genotyping data the use of sCVPP is natural albeit at the cost of the significant assumption that there is at most 1 causal variant (per trait) in any genetic locus under consideration. One benefit is that sCVPP can, through simple summation, be combined within a given genetic locus, to estimate posterior probabilities that a given feature overlaps a causal variant. This property facilitates the integration of GWAS data at a single basepair resolution, within the larger *Hin*dIII regions underlying PCHi-C contact maps. Another attractive feature of sCVPP is that they are computed from a joint model of SNPs across an LD block (Wellcome Trust Case Control Consortium et al., 2012) and thus naturally adjust for LD between variants within that block.

3.4.2 Annotation of coding variants

Whilst PCHi-C data has the prospect of linking non-coding variation to target genes, any systematic method should take also into account other functional annotation classes that occur in the coding space. To do this I annotated all SNPs within dbSNP 138 (Sherry et al., 2001) using VEP (McLaren et al., 2016) employing gene annotations from Ensembl v75 (Yates et al., 2016). I then filtered these annotations so that I obtained the set of SNPs that overlapped protein coding genes. Whilst there is some literature that implies that coding variation may have regulatory potential (Stergachis et al., 2013), this has proved controversial (Xing and He, 2015) and I chose to assign function to such variants exclusively within the gene within which they occurred. I made no distinction between classes of coding variation as although non-synonymous coding variants by definition alter protein sequence, there is evidence that synonymous variants can effect profound changes through, for example the introduction of cryptic splice sites (Rice et al., 2013). Thus, the final set of coding SNPs (cSNPs) included all exonic variants in protein coding genes annotated by VEP/Ensembl v75.

3.4.3 Annotation of PCHi-C 'blindspot' ('Virtual Promoter')

A limitation of the PCHi-C method is that interactions between a captured fragment and its adjacent *Hin*dIII fragments are not able to be reliably called, as signal is overwhelmed by random Brownian effects (Cairns et al., 2016). In order to

capture these I created a virtual 'promoter' fragment (VPF) by merging each capture fragment (bait) with those immediately 5' and 3' (Figure 3.4b).

3.4.4 Annotation of PIRs

I defined PIRs for a gene as those contributing an interaction exceeding a CHiCAGO threshold of 5 as recommended by Cairns et al. (2016). In the case where a gene has multiple baits, due to the presence of multiple TSS, it is possible for the same PIR to be linked to a gene more than once. In such an instance, I only considered the highest scoring PIR within a given tissue when applying CHiCAGO thresholding.

3.4.5 COGS method description

Consider a region of the genome, r that contains n_r variants $v_{i,i\in r}$ where i indexes SNPs across the genome and the notation $i \in r$ means SNP_i is located within region r. A region can contain intervals, that represent one of three categories of genomic features previously described: coding SNPs (Section 3.4.2), VPF (Section 3.4.3) and PCHi-C PIRs (Section 3.4.4). The posterior probability for any feature or set of features, f_r, overlapping r to contain the causal variant under a single causal variant assumption is

$$\sum_{i:v_i \in f_r} \text{sCVPP}_i, \tag{3.1}$$

where sCVPP_i is the single causal variant posterior probability for the i^{th} variant. Across m regions $r_1, r_2, \ldots, r_m$ there may be at most m causal variants, and I calculate the posterior probability that sets of features spread across these regions contain at least one causal variant as one minus the posterior probability there is no causal variant in any of them, on the assumption that there is complete independence between the regions,

$$\text{COGS}_g = 1 - \prod_{j=1}^{m} \left(1 - \sum_{i:v_i \in f_{r_j}} \text{sCVPP}_i \right). \tag{3.2}$$

I select the regions such that they are approximately LD independent, and so the assumption of independence may not strictly hold, for this reason the computed value is not a true posterior probability and I use the term COGS score instead. When the set of features are defined as all those linked to a specific gene g, I denote this COGS_g.

The above framework makes the assumption that features are disjoint (i.e. non-overlapping), however in practice overlap exists. For example variants in the coding space will often overlap VPFs due to their proximity to gene bodies (Figure 3.4b). To overcome this, prior to computing the COGS score, I compute the union of all features across all categories, that is the smallest set of features required to cover all categories, and use this as the input into the framework.

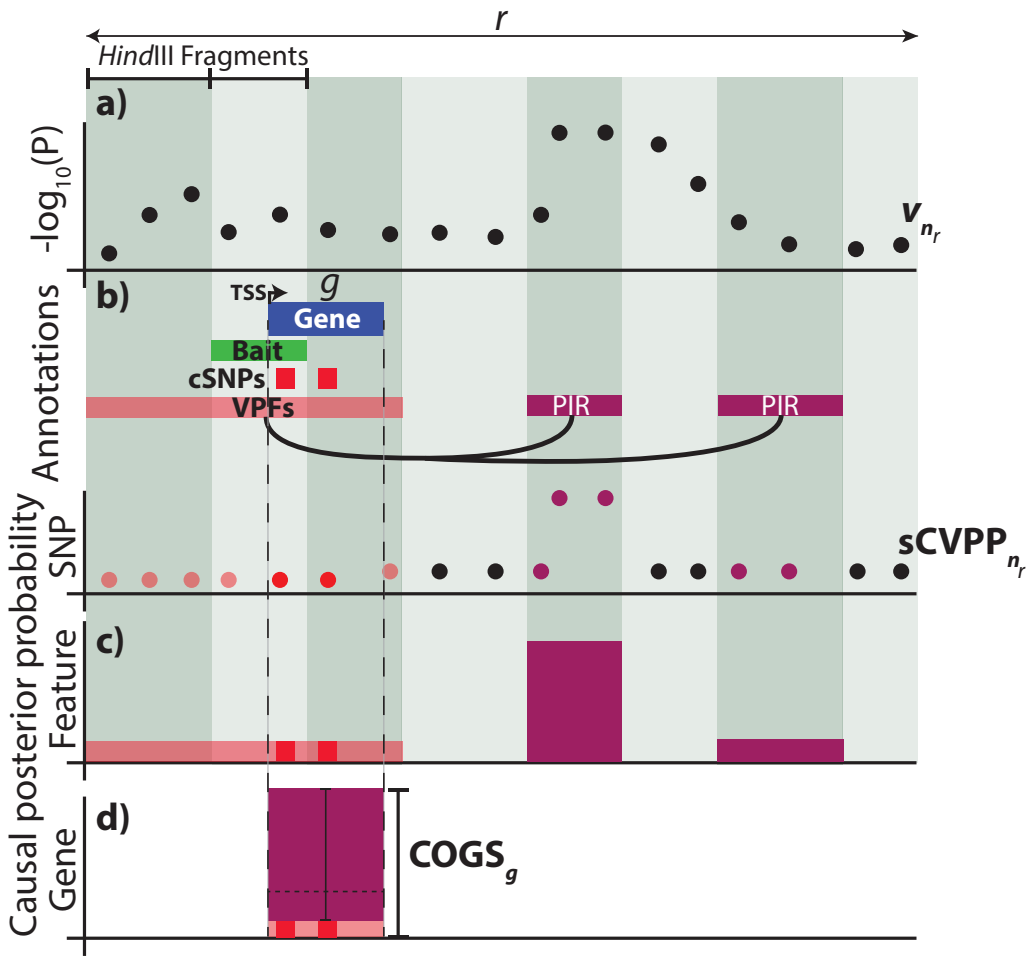


Fig. 3.4 A schematic illustration of the COGS (**C**apture Hi-C **O**mnibus **G**ene **S**core) method. **a)** GWAS summary statistics, for region r, are converted to single causal variant posterior probabilities (sCVPP$_{n_r}$). **b**) These are intersected with three feature sets; VPFs (pink), cSNPs (red) and PIRs (purple) linking to the focal gene, g (blue). **c)** Causal feature posterior probabilities are combined through summation to calculate posterior probabilities for a given feature set to be causal (Equation 3.1). **d)** Finally these are summed to compute an overall COGS score for g to be causal. In most cases features span multiple regions, which assuming independence are combined to generate an overall COGS score (Equation 3.2).

3.5 Application of COGS to a GWAS compendium

Initially, in order to simplify the assessment of COGS performance, I considered the non-redundant set of PIRs linking to a gene, *g*, across all 17 primary tissues, along with VPF and cSNPs. It is worth restating that variants overlapping the HLA region were removed prior to running COGS, due to both its extended LD structure and known strong and complex association with immune-mediate disease, such that a single causal variant assumption would not be defensible.

3.5.1 Overall COGS scores for 31 traits

I used COGS to integrate this aggregated contact map with the compendium of PMI imputed GWAS summary statistics for each of the traits detailed in chapter 2 (Table 2.1). Across all traits, the mean count of protein coding genes for which COGS scores were generated was 16,910; differences in counts between traits are expected due to variability in input GWAS summary statistic coverage. Overall, the majority of gene scores were close to zero, with 99% of genes having a score less than 0.05 (Figure 3.5).

I created a set of 'high scoring' genes for each trait by selecting those with an overall COGS gene scores greater than 0.5. In total, across all traits, 2,604 unique protein coding genes were prioritised, with a median of 112 genes prioritised per trait (Figure 3.6a).

Height had the most genes prioritised, and INS the least. It is established that the number of GWAS associations discovered relates to sample size, therefore I looked to see whether there was evidence for a relationship between sample size and the number of genes prioritised (Figure 3.6B). I observed no correlation between sample size and prioritised gene count (Spearman's $\rho \approx -0.09, P \approx 0.6$). This is not unexpected as the number of genes prioritised will also be affected by many other factors, including trait heritability and the relevance of primary blood cell specific PCHi-C data used.

3.5.2 Are prioritised genes biologically relevant?

Due to the large set of genes prioritised I was motivated to perform a systematic analysis across all traits in order to provide support for their biological relevance. One approach, GSEA, can be a useful indicator as to whether a set of empirically described genes, such as those prioritised by COGS, are enriched for a particular biological function. I performed a GSEA of prioritised COGS genes over all traits

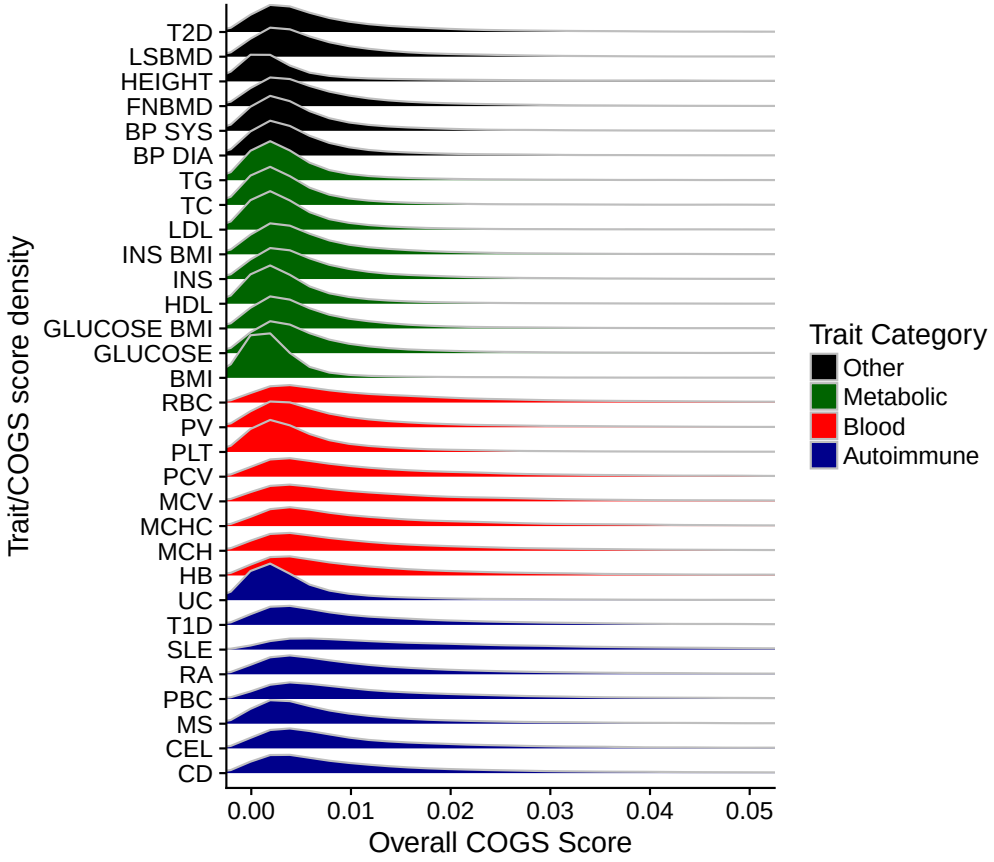


Fig. 3.5 A density plot of COGS scores across all traits. For clarity the horizontal axis is truncated at 0.05, although the observations extend to 1.

using the Reactome (Fabregat et al., 2016) resource as a source of curated gene sets.

I downloaded the Reactome genesets in `gmt` format from MolSigDB (v6.2) (Liberzon et al., 2015; Subramanian et al., 2005), filtering each gene set so as to only include those genes for which a COGS score could be computed. I used Fisher's test in order to assess whether there was evidence for enrichment of COGS prioritised genes (Overall COGS score > 0.5) compared to background (Overall COGS score ≤ 0.5) for each of 674 genesets. I selected as significant those with a Benjamini-Hochberg FDR $< 5\%$, calculated across all gene sets and baits (Figure 3.7).

As expected, genes prioritised for autoimmune diseases (blue, Figure 3.7) were enriched in inflammation and immune-response-related gene sets, such as T cell receptor signaling, whereas genes prioritised for metabolic traits (green, Figure 3.7) were preferentially enriched for lipid transport and metabolism.

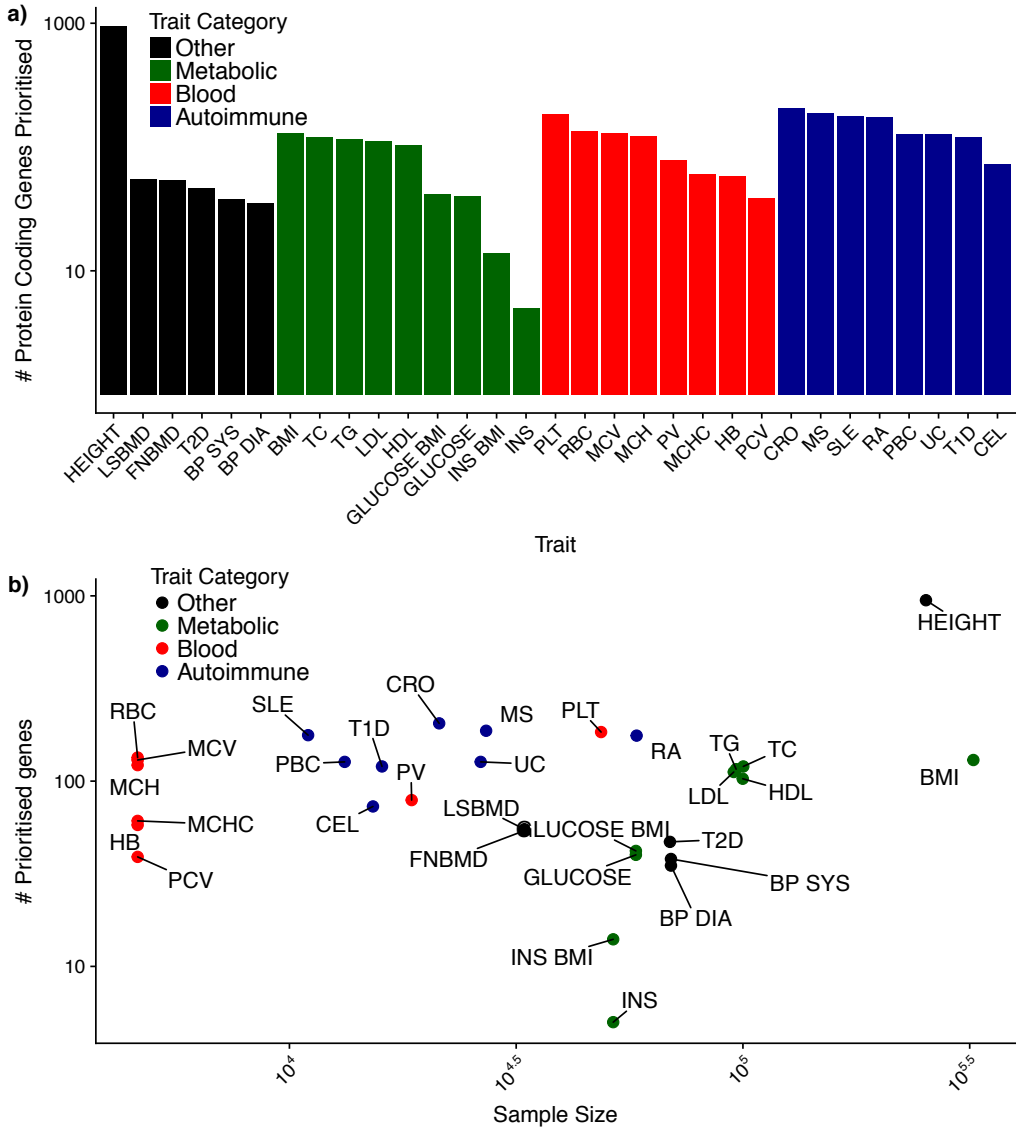


Fig. 3.6 **a)** Counts of prioritised COGS protein coding genes (Overall gene score > 0.5) for 31 GWAS traits. **b)** GWAS sample size vs number of genes prioritised, axis are $\log_{10}$ scaled.

As described in the previous chapter, robust GSEA with structured data such as those encountered when integrating PCHi-C and GWAS is challenging and thus such analysis is limited. Firstly, there will be correlation between COGS scores due to LD and because PIRs can be shared between genes. Secondly, tissues assayed by PCHi-C are derived from haematopoietic lineages, thus biasing enrichment towards genesets relevant to these tissues. Nevertheless, such an analysis does provide limited support that genes prioritised by COGS are broadly of biological relevance to the trait of interest.

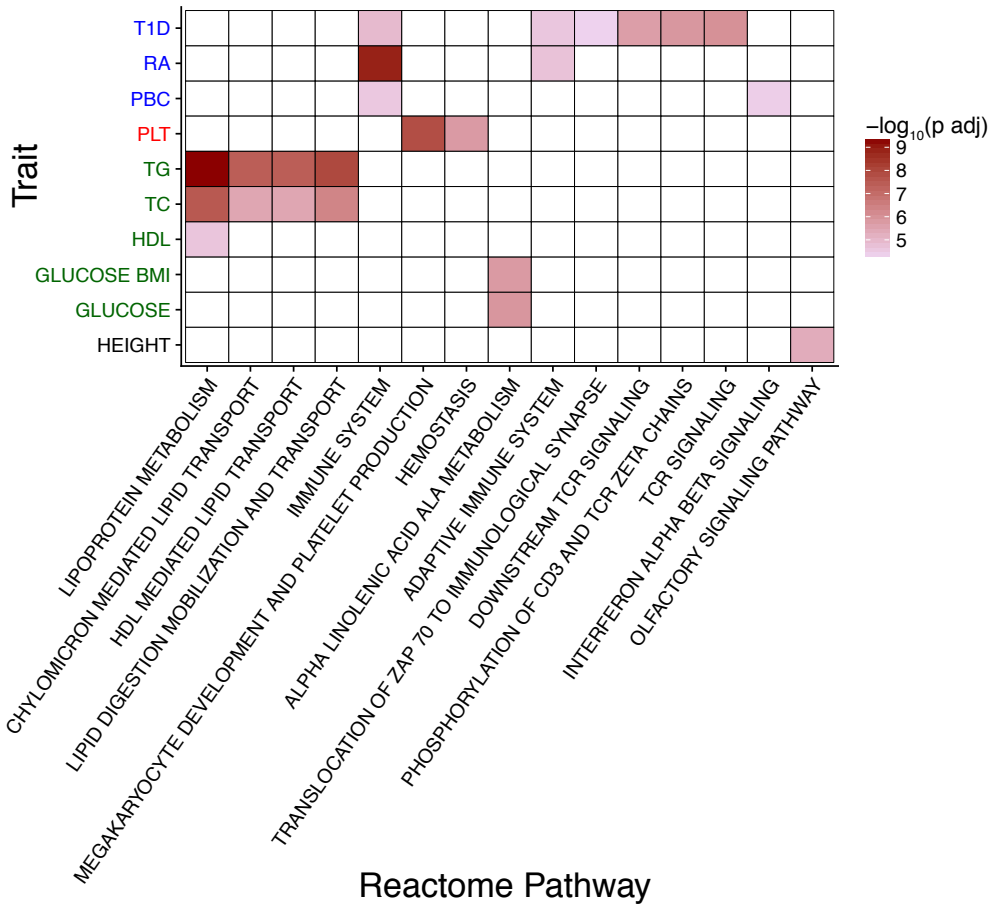


Fig. 3.7 Trait specific enrichment of COGS prioritised genes within Reactome gene sets. For clarity traits and gene sets for which no significant enrichment (FDR<5%) was observed have been omitted. To facilitate interpretation, non significant comparisons are shown in white. Traits are coloured according to categories defined in Figure 3.5, blue - IMD, red - blood, green - metabolic and black - other.

3.6 Impact of different gene-variant linking methods on COGS performance

COGS is a novel method and therefore it is important to compare its performance under the scenario of different inputs, including PCHi-C, for linking variants to genes. Such an assessment is challenging, however, as there are no 'gold standard' or comprehensive lists of causal disease genes. Indeed, only a handful genes exist for which functional studies can provide strong evidence for their causality (Cho and Feldman, 2015; Dendrou et al., 2016; Gregory et al., 2012; Klein et al., 2005).

There are even fewer that have been characterised convincingly in a non coding context (Claussnitzer et al., 2015; Dendrou et al., 2009).

3.6.1 A framework for comparing gene-variant linking methods using GWAS summary statistics

I examined three inputs for linking variants to genes as follows:-

Proximity based intervals: For each protein coding gene, I isolated the set of *Hin*dIII fragments overlapping the relevant TSS and that were captured on the PCHi-C platform. I extended each fragment 5′ and 3′ by 0.5Mb, resulting in a set of genomic intervals of approximately 1Mb in size. Whilst arbitrary such a method will capture most variants in high LD around the target gene. When multiple TSS for the same gene were captured in different fragments, I took the union of such intervals in order to obtain one interval for each gene.

Topologically associated domain (TAD) based intervals: For each protein coding gene and tissue context (Table 3.2), I defined a TAD based proximity region as the union of TADs physically overlapping a *Hin*dIII containing a relevant TSS captured on the PCHi-C platform. TADs were obtained from conventional Hi-C analysis of eight primary cell-types by Csilla Varnai, Michiel Thiecke and Mikhail Spivakov (Table 3.2).

PCHi-C based intervals: For each protein coding gene and tissue I defined a PCHi-C interval as the union of all PIRs (CHiCAGO score > 5) across tissues (for which TAD boundary data was available) called for any PCHi-C captured protein coding promoter *Hin*dIII fragment.

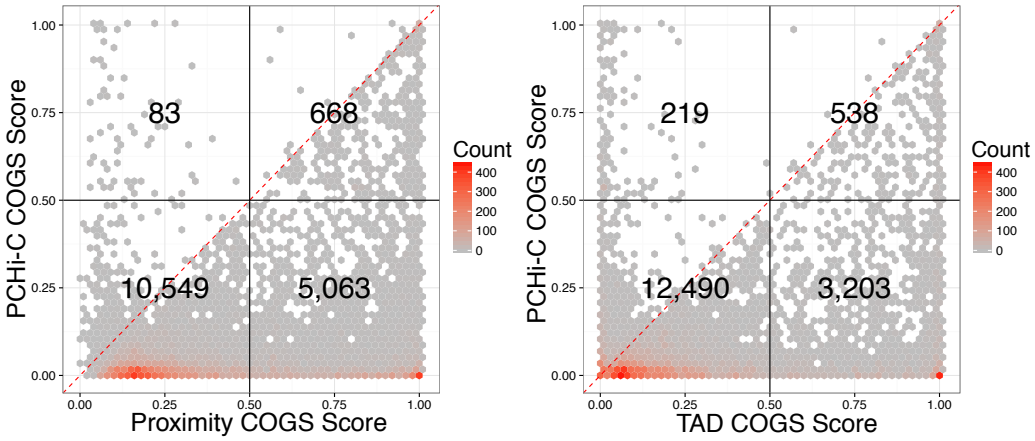
For each of the sets of intervals described above I generated COGS scores across the eight IMDs included in the GWAS compendium (multiple sclerosis, coeliac disease, type 1 diabetes, Crohn's disease, primary biliary cirrhosis, ulcerative colitis, SLE and rheumatoid arthritis), using PMI data sets from which I had masked cSNPs and those SNPs mapping to the HLA region. PCHi-C and TAD intervals generated tissue specific scores, and to facilitate their comparison with proximity derived scores, I took forward the maximum score for a given gene across all eight tissues considered.

Tissue	TAD Coverage (Gb)
Erythroblasts	1.53
Macrophages	1.68
Megakaryocytes	1.59
Monocytes	1.48
Naive B cells	1.51
Naive $CD4^+$ T cells	1.40
Naive $CD8^+$ T cells	1.51
Neutrophils	1.27

Table 3.2 Topologically associated domain coverage across eight cell types elucidated from classical Hi-C analysis

3.6.2 Comparison of PCHi-C-, proximity- and TAD-based COGS scores

Figure 3.8 shows COGS scores derived from TAD and proximal derived intervals compared to those from PCHi-C. I categorised each gene according to whether it was prioritised (interval method specific COGS score > 0.5) by both, a single or neither method. For example, Figure 3.8(a), shows the comparison of proximity (horizontal axis) and PCHi-C derived COGS scores, here the number in the top right quadrant, 668, is the number of genes prioritised across all eight traits by both methods.



(a) (b)

Fig. 3.8 Comparison of autoimmune PCHi-C COGS scores and (a) proximity COGS score from assigning variants to genes within 0.5 Mb of gene promoters (b) Hi-C derived TAD COGS scores, using seven Cell types (Erythroblasts, Macrophages, Monocytes, Naive B cells, Naive $CD4^+$ T cells, Naive $CD8^+$ T cells and Neutrophils). Counts of genes in each quadrant are shown, grey to red colour gradient indicates gene density.

In general PCHi-C COGS prioritised genes sets were smaller than then those from the other methods. The naive proximity based COGS score appeared to be least explicit prioritising 5,731 genes in total, however it had the greatest overlap with PCHi-C COGS scores. On comparing proximity based and PCHi-C derived scores, 83 (Figure 3.8(a)) genes were prioritised exclusively by PCHi-C COGS, suggesting that 12% of PCHi-C prioritised genes are related to interactions greater than 0.5 Mb. Alternatively, when comparing PCHi-C and TAD derived COGS scores, 219 (Figure 3.8(b)) genes were prioritised exclusively by PCHi-C COGS, indicating that approximately 30% of PCHi-C prioritised genes relate to interactions that span TAD boundaries. To explore this further I looked at the distance between genes prioritised by TAD and PCHi-C inputs (COGS score > 0.5) and the distribution of distances to the closest TAD boundary (Figure 3.9). Genes prioritised solely by PCHi-C COGS seemed to be closer to TAD boundaries compared to those prioritised specifically by the TAD method. However, the distribution of genes prioritised by both methods is similar indicating that this phenomenon is potentially due to the imprecise nature of TAD boundary definition, rather than a true underlying biological phenomenon.

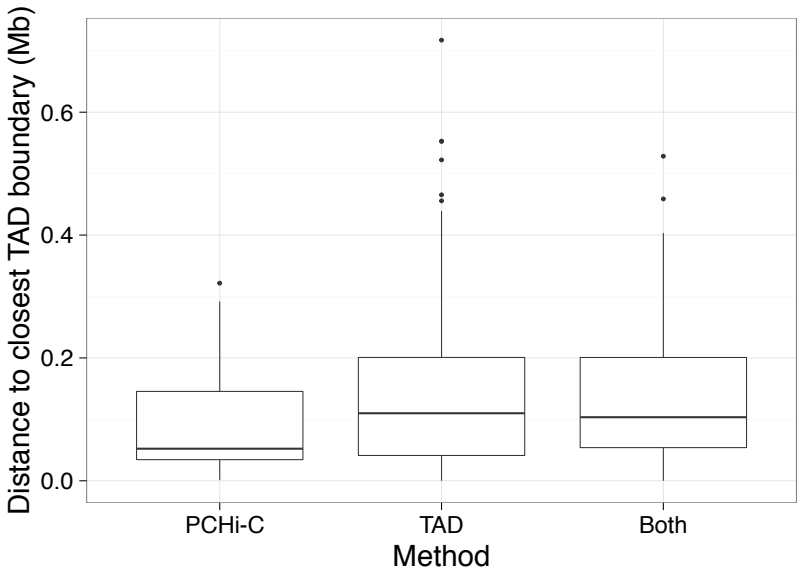


Fig. 3.9 Box plot showing the distribution of distances between baits and TAD boundaries for significant (score > 0.5) genes. 'Both' indicates that gene was significant using TAD and PCHi-C methods.

In summary, COGS using PCHi-C as input selects genes not found by other inputs whilst simultaneously prioritising far fewer genes.

3.6.3 PCHi-C prioritised genes are more likely to be differentially expressed in disease patients

In addition to quantifying the relative size of prioritised gene sets, I wanted to assess evidence for biologically plausibility. I used data from a study (Peters et al., 2016) of gene expression data across a range of relevant tissues for individuals with active inflammatory bowel disease (IBD) and healthy controls, hypothesising that biological plausibility would be reflected in enrichment of genes differentially expressed between patients and controls within a prioritised gene set.

To generate a set of differentially expressed genes, I downloaded the data from ArrayExpress (Kolesnikov et al., 2015) (E-MTAB-3554). The data set consists of PEER (Stegle et al., 2012) normalised microarray expression values across 49 patients with Crohn's disease, 42 with ulcerative colitis and 43 healthy controls, across sorted $CD4^+$ T cells, $CD8^+$ T cells, B cells, Monocytes, and Neutrophils. I modified an R script from Chris Wallace to compute differential expression using `limma` (Ritchie et al., 2015) with a null hypothesis that expression for a given gene was the same across all three groups within a tissue. As PCHi-C and TAD COGS scores calculated for the previous comparison are derived by combining over cell types I selected the union of genes differentially expressed in at least one cell type.

To generate a set of prioritised genes for the IBD component diseases, ulcerative colitis and Crohn's disease, I used input GWAS statistics from Anderson et al. (2011) and Franke et al. (2010) with coding and HLA variants removed. I computed PCHi-C, TAD and proximal gene scores using only cell-types matching those in Peters et al. (2016).

I used Fisher's test to examine enrichment for prioritised genes, for each of the three methods, for genes showing evidence of differential expression between healthy, and Crohn's disease (CD) or ulcerative colitis (UC) at a 5% false discovery rate. I found PCHi-C genes were enriched for differentially expressed genes for UC ($P = 0.002$) and CD ($P = 0.04$). I found no evidence of enrichment in any data sets using proximal and TAD methods (Figure 3.10).

Whilst such an analysis is not conclusive, it provides further evidence that integrating PCHi-C data with GWAS through COGS can prioritise biologically relevant genes, and that this is in some sense more informative, evidenced by smaller prioritised gene lists and higher enrichment in a biologically relevant expression dataset than than proximity or TAD based COGS scores.

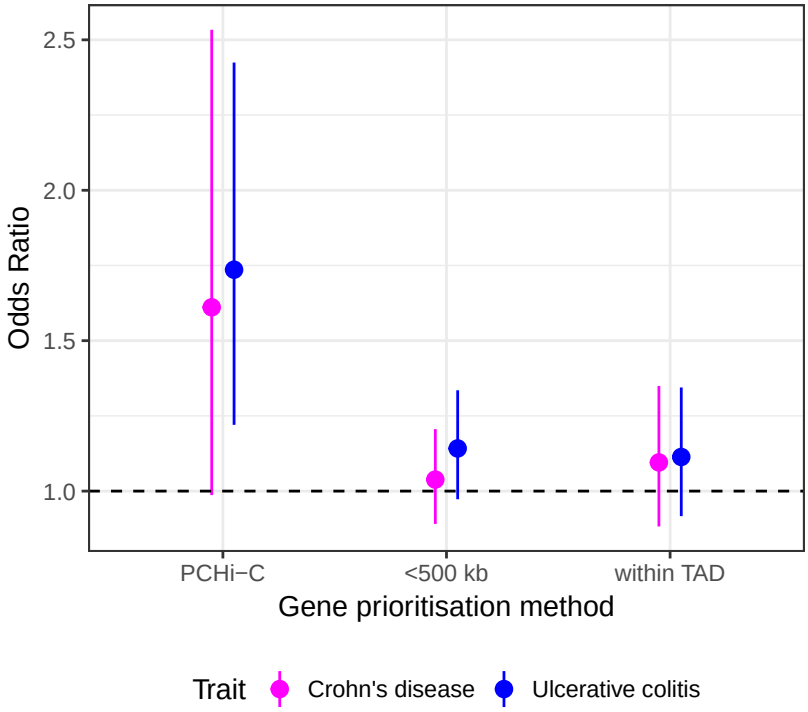


Fig. 3.10 Enrichment of prioritised genes in ulcerative colitis and Crohn's disease differentially expressed genes from Peters et al. (2016).

COGS identification of biologically relevant genes in Crohn's disease

The intersect between differentially expressed genes and COGS prioritised genes identified 67 genes (Table B.1). Whilst there are many candidate genes in this list, an example on chromosome 3 for ulcerative colitis provides an illustration of the potential for hypothesis generation by integrative analysis of PCHi-C interaction maps (Figure 3.11). COGS prioritises *BCL6* which has been shown to have potent effects on Th9 cell development and IL-9 secretion, both important modulators of inflammation (Bassil et al., 2014). This prioritisation results from putative causal variants suggested by Anderson et al. (2011) and more recently de Lange et al. (2017) for Crohn's disease (CD) that reside in intron 8 of *LPP*. Whilst previous studies have implicated a causal role for *LPP*, PCHi-C results from lymphoid and myeloid tissues show interactions between this region and the *BCL6* promoter (Figure 3.11).

Expression profiles from Peters et al. (2016) show that whilst *LPP* is not significantly differentially expressed, in monocytes, between CD controls and healthy volunteers (log(Fold Change) = 0.05, $P_{adj} = 0.46$) *BCL6* is (log(Fold Change) = 0.22, $P_{adj} = 0.0018$), providing further support for its prioritisation.

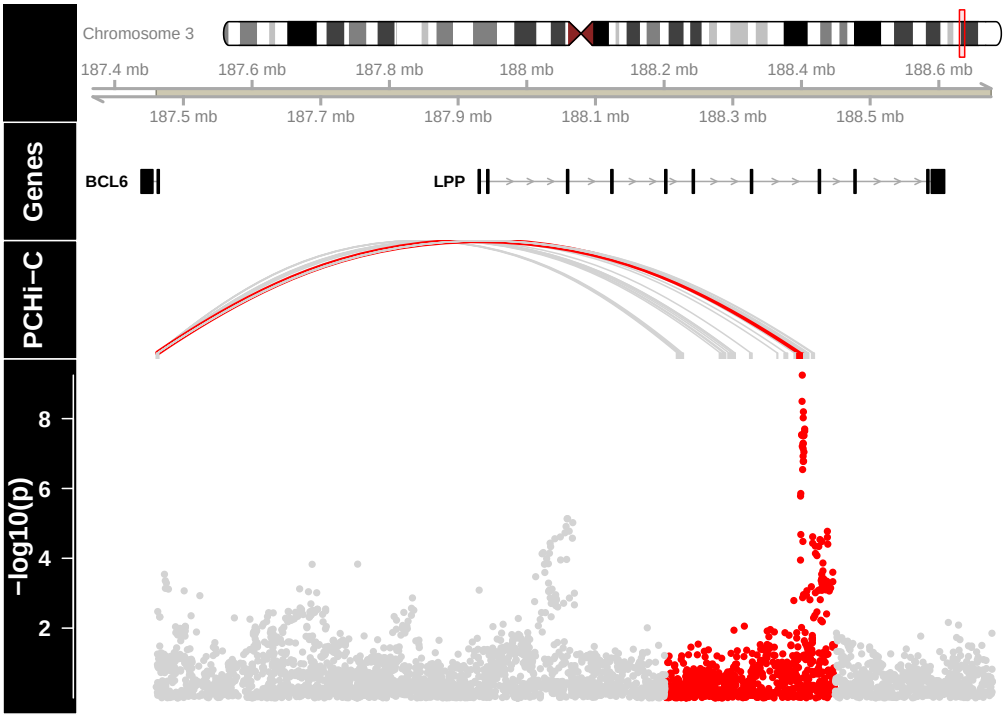


Fig. 3.11 Example of COGS prioritised gene on chromosome 3, '$-\log_{10}(P)$' stanza values are taken from de Lange et al. (2017) GWAS of Crohn's disease, the recombination block containing the index SNP rs56116661 is shown in red. 'PCHi-C' stanza shows significant PCHi-C interactions in Macrophages between the recombination block containing the index SNP and the promoter of *BCL6* gene. The interaction marked in red overlaps rs56116661, with > 0.5 posterior probability of being the causal variant according to de Lange et al. (2017).

3.6.4 Overlap of COGS prioritised genes with eQTLs

Another line of evidence supporting the functional significance of PCHi-C interactions is the enrichment of eQTL signals in PIRs where the target gene is shared between the two methods. Roman Kreutzhuber under the supervision of Oliver Stegle had observed such a tissue specific (Fairfax et al., 2012) enrichment (Javierre et al., 2016), and given the enrichment of GWAS signals in tissue specific PIRs that I had observed, it was natural to consider the extent of overlap between COGS prioritised genes and eQTL signals.

I took forward COGS prioritised genes in systemic lupus erythematosus (SLE) (Bentham et al., 2015) and rheumatoid arthritis (RA) (Okada et al., 2014) data sets, for which full imputed summary statistics were available. Out of 456 genes that were prioritised for either trait 136 had eQTLs of which four genes (*BLK*, *RASGRP1*, *SUOX*, and *GIN1*) showed evidence for possible co-localisation

in RA and two genes (*BLK* and *SLC15A4*) in SLE. Additionally the genes prioritised for RA included 5/9 candidates (*C8Orf13*, *BLK*, *TRAF1*, *FADS2* and *SYNGR1*) that were identified in a study that combined whole blood eQTLs with the same RA GWAS data by Mendelian randomisation (Zhu et al., 2016). The relatively large number of GWAS prioritised genes without eQTL support agrees with previous reports of limited overlap of disease variants with eQTL datasets (Guo et al., 2015; Huang et al., 2017).

3.7 Comparison of PCHi-C COGS scores between tissues

Analyses in Chapter 2 have shown that tissue specific PCHi-C interactions show preferential enrichment for GWAS signals. This motivated me to extend the COGS framework, such that it would be suitable for jointly prioritising causal genes *and* tissue contexts. In Section 3.4.5, I ignored tissue specificity to simplify initial method development and assessment. However, by restricting the sets of features available to COGS it is possible for it to operate in a tissue specific manner.

In section 3.4.5 I showed how, under the assumption of at most a single causal variant within a genomic region r, the posterior probability for a fragment to contain a causal variant is the sum of sCVPP over SNPs overlapping that fragment. The fact that the coordinates of each *Hin*dIII fragment are known *a priori* allows fragment posterior probabilities to be computed and stored for a given trait prior to computing a specific gene's COGS score. However, given that the set of r_m regions are selected, not on the basis of alignment with *Hin*dIII fragments, but for approximate LD independence, adjustment is necessary. This entails the identification and splitting of all fragments overlapping approximate LD independent region boundaries, which I term 'LD aware' *Hin*dIII fragments (Figure 3.12a). Coding variation presents an additional complication, as it needs to be assessed on a gene-by-gene basis (Section 3.4.2), however again all cSNPs are known *a priori*, as such it is simple to compute 'LD aware' *Hin*dIII fragments with all cSNPs removed (Figure 3.12b).

With 'LD aware' *Hin*dIII fragment posterior probabilities for a given trait precomputed the generation of tissue specific COGS scores is greatly facilitated, as it is a simple operation of selecting fragment posterior probabilities corresponding to a gene and tissue of interest for a given region, adding in cSNPs as for the focal gene, and then combining across regions using Equation 3.2 (Figure 3.12c).

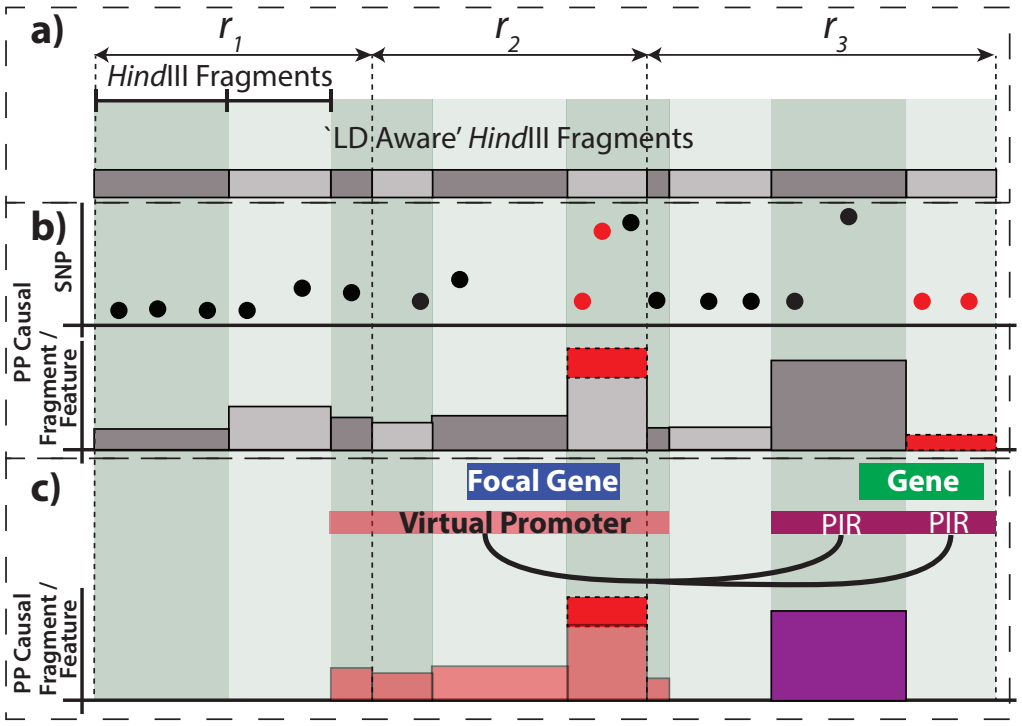


Fig. 3.12 Computation of feature specific COGS scores. **a)** The locus shown consists of three approximately LD independent regions, r_1 to r_3, which are intersected with *Hind*III fragments to create 'LD aware' fragments (alternating grey) with all coding variants (red) removed. **b)** sCVPP computed using the same LD independent regions are intersected with these 'LD aware' *Hind*III fragments to obtain causal fragment posterior probabilities, which only needs to occur once per trait. **c)** These fragment posterior probabilities can then be annotated with any combination of VPF (pink), tissue specific PIRs (purple) and coding variants (red) from which a feature specific COGS score can be calculated for the focal gene (blue). In this example, coding variation in an alternative gene (green) is not considered even though there is overlap with a focal gene PIR.

Such a scheme is fast and flexible, for example, say we wished to compute a COGS gene score for lymphoid cells *without* taking into account VPF and cSNP evidence. In this case I would select the union of PIRs across nine cell types (nB, tB, FetT, aCD4, naCD4, tCD4, nCD8, nCD4, tCD8 - Figure 3.13) taking forward the fragment posterior probabilities of the non redundant set of PIRs with CHiCAGO scores greater than five. As for each trait fragment posterior probabilities (excluding cSNPs) are stored, the computation of the COGS score requires a simple lookup of the relevant fragments, which can then be combined using Equation 3.2.

3.7.1 A heuristic approach to prioritising sets of tissues

A naive approach to systematically analyse all combinations of feature sets would require $2^{19} - 1$ combinations of features for each gene, as there are 17 tissue feature sets in addition to coding and VPF features. This motivated me to develop a heuristic approach that utilised the overall relationships between PCHi-C tissues learned from the data in Chapter 2 through the application of PCA and hierarchical clustering (Figure 2.4).

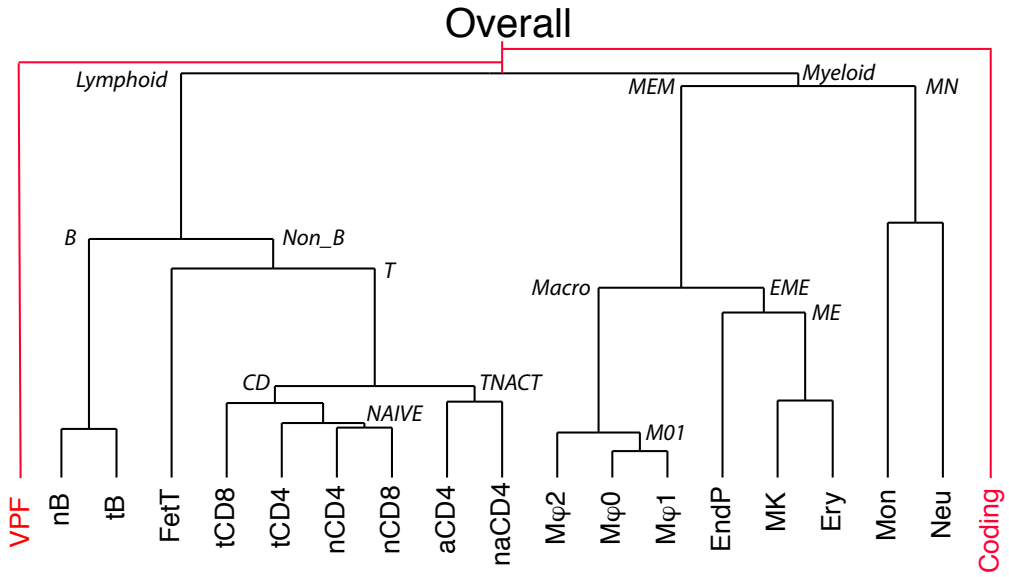


Fig. 3.13 Hierarchical clustering of PCHi-C by CHiCAGO score profile that broadly reconstitutes the Haematopoietic tree modified from Chapter 2 to include two non-PCHi-C nodes (red) **VPF** - 'Virtual' promoter fragments and **Coding** - focal protein coding variants. PCHi-C tissues (black) are labelled as follows: **Lymphoid**; nB - naive B cells, tB - total B cells, FetT - Fetal Thymus, aCD4 - activated $CD4^+$ T cells, naCD4 - non-activated $CD4^+$ T cells, tCD4 - total $CD4^+$ T cells, nCD8 - naive $CD8^+$ T cells, nCD4 - naive $CD4^+$ T cells, tCD8 - total $CD8^+$ T cells. **Myeloid**; Mon - Monocytes, Neu - Neutrophils, Mϕ2 - M2 Macrophages, Mϕ1 - M1 Macrophages, Mϕ0 - M0 Macrophages, EndP - Endothelial Precursor cells, MK - Megakaryocytes, Ery - Erythroblasts. The Overall label illustrates the COGS scores initially computed in Section 3.4.5 and combines both PCHi-C (black) non PCHi-C (red) features. Nodes, representing collections of clustered tissues at different hierarchies, are labelled in italics.

In order to incorporate the non PCHi-C featuresets I added to the root of this tree two additional nodes representing the possibility of a variant to act through a coding variant or to be present in a VPF (Section 3.4.3), for which tissue assignment is not possible (marked in red in Figure 3.13).

For a given GWAS I use COGS to compute a score for each binary decision point or node. For simplicity consider a single region containing a focal gene for which COGS scores derived from two feature sets F and F' are to be compared. The ratio of these scores is a ratio of likelihoods under different hypotheses or a Bayes factor (Kass and Raftery, 1995). COGS scores, however incorporate information across multiple regions that are assumed to be independent of one another. A COGS score ratio therefore incorporates products of (possibly) dependent probabilities. In order to disambiguate and highlight this I therefore use the term them psuedo-Bayes Factor (pBF) interpreting these pBF in a similar manner to Kass and Raftery (1995), such that they represent 'a summary of the evidence provided by the data in favour of one scientific theory, represented by a statistical model, as opposed to another'.

For example (Figure 3.14) at the root or 'Overall' node I compute two COGS gene scores; one considering only coding variants and the other considering only non-coding variants, that is those located in target gene VPFs or overlapping any target gene PIR across all 17 primary blood cell types. The ratio of COGS scores incorporating these different feature sets therefore reflects the strength of evidence for a single putative causal variant exerting its effect through an alteration of protein coding sequence, as opposed to a broader non-coding space alternative. If the evidence for these models is balanced, such that there is equal support for either, then the pBF ≈ 1, and the method returns an 'Overall' hypothesis label indicating that a coding and non-coding mechanism are statistically not discriminable (Figure 3.14). Alternatively, if the data supports one of the binary hypotheses over the other then that hypothesis is returned. I again refer to Kass and Raftery (1995) for guidance on what magnitude of pBF provides a suitable support threshold, employing pBF $< \frac{1}{3}$ or pBF > 3, to support a given binary choice and pBF $\geq \frac{1}{3}$ or pBF ≤ 3 as 'balanced' (inconclusive) evidence to discriminate between hypotheses.

I apply these tests in a hierarchical manner 'stepping' down the tree when evidence is strong enough to support a decision, or sticking otherwise. For example if a 'Non-coding' hypothesis is favoured over 'coding' we move to the 'Non-coding' node (Figure 3.14) and compute a pBF for the next binary choice, whether the data favours a causal variant operating through a virtual 'Promoter' region or through a PCHi-C identified chromatin looping or 'Interaction' event. This biologically structured and hierarchical examination of the hypothesis space continues until either $\frac{1}{3} \leq$ pBF ≤ 3, such that no decision is favoured or we reach a terminal 'leaf' node for which further categorisation is not possible (Figure 3.14).

This hierarchical COGS analysis results in COGS scores at each node, and I introduce the term lCOGS scores to disambiguate between COGS scores at terminal 'Leaf' nodes from their more general counterparts.

3.7.2 Tissue specific COGS gene prioritisation across 8 immune-mediated diseases

I applied the hierarchical COGS method described to the compendium of GWAS traits for protein coding genes where the overall COGS score (including all tissue PIRs, coding variants and VPF) exceeded 0.5 in order for them to be comparable with results described in Section 3.5.1. Due to haematopoietic focus of the PCHi-C interaction data I restricted my analysis to only blood and immune based traits, as for metabolic and anthropomorphic traits haematopoietic contexts are unlikely to be truly relevant for discrimination between cell types. I arranged the counts of genes at each node for each trait into a matrix, and performed hierarchical clustering using the 'average' linkage method on a $\log_{10}(\text{count}+1)$ transformed matrix in order to accommodate the large range $(0-694)$ of values encountered (Figure 3.15).

Generally the number of genes able to be assigned to specific cell types or leaf nodes was low (IQR $0-2$, mean 1.75), with mean gene counts ranging from 0.3 in total $CD8^+$ T cells to 9.7 in neutrophils. In comparison non-leaf nodes counts were higher (IQR $0-4$, mean 8.1) with a large proportion driven by non tissue-specific interactions. This extensive sharing of PIRs across cell-types, could be due to shared underlying transcriptional programs between related tissue types or due to the limitations in PCHi-C resolution, with disparate but tissue specific enhancers sharing a common PIR *Hin*dIII fragment. Secondly, in regions of high LD sCVPP will be shared across many variants, reducing resolution, and thus making it more likely that collateral PIRs in disparate tissues will contribute to the final COGS score. Such effects will be compounded by the genetic heterogeneity and study design of the traits included in the GWAS compendium. With the current approach it is impossible to disentangle the effect of power, PMI and heritability from these counts in order to make robust biological inference.

With these caveats in mind, there are notable features that arise from the hierarchical clustering. The foremost is that generally immune-mediated traits (blue in Figure 3.15), with the exception of coeliac disease (CEL) appear to cluster. On explanation might be that CEL may be affected by limited genotype coverage, and thus a greater reliance on PMI, which is compounded by sample

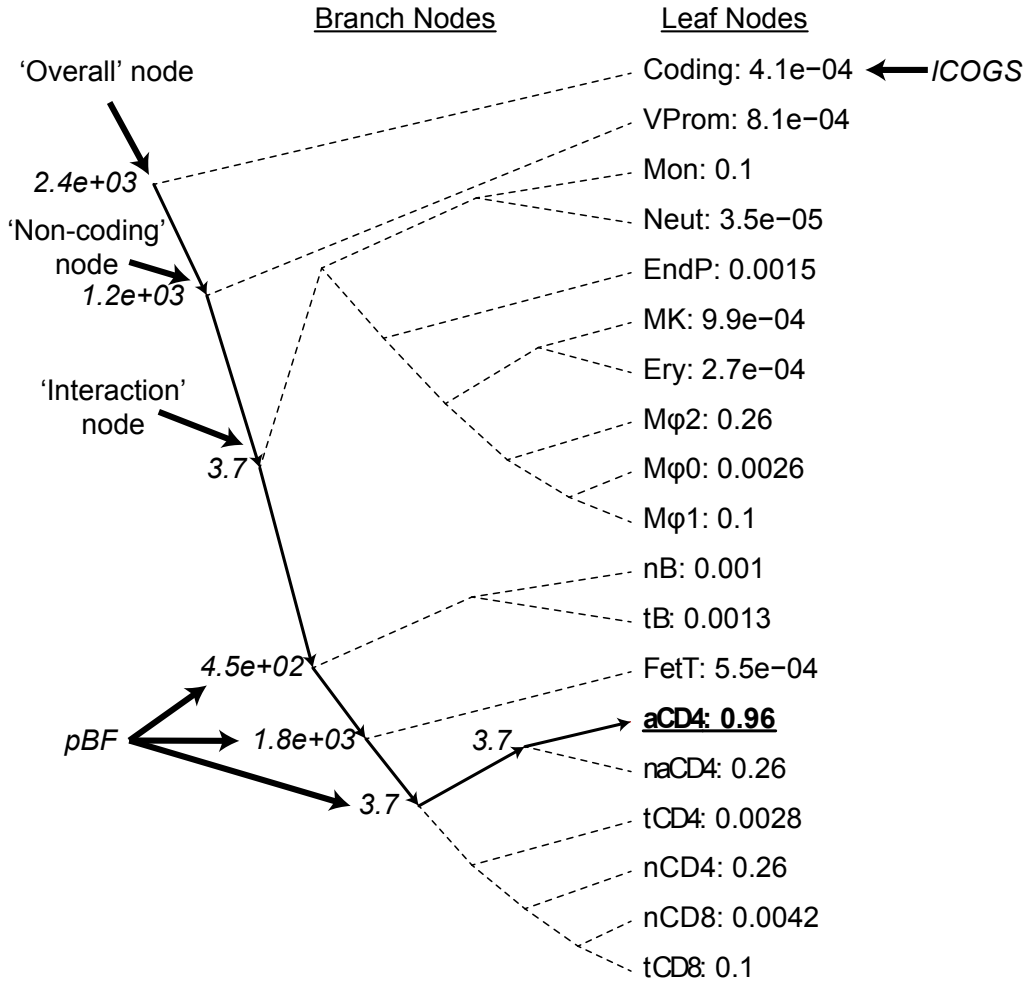


Fig. 3.14 A cladogram illustrating the hierarchical method for assigning putative category labels to genes based on COGS scores for the *AHR* gene in the context of RA GWAS summary statistics (Okada et al., 2014). The solid edges denote the path taken by COGS through the binary decision tree. Each selected node is labelled with an italicised pseudo Bayes factor. Leaf nodes representing specific tissues or functional categories are labelled as follows; nB - naive B cells, tB - total B cells, FetT - Fetal Thymus, aCD4 - activated $CD4^+$ T cells, naCD4 - non-activated $CD4^+$ T cells, tCD4 - total $CD4^+$ T cells, nCD8 - naive $CD8^+$ T cells, nCD4 - naive $CD4^+$ T cells, tCD8 - total $CD8^+$ T cells, Mon - Monocytes, Neu - Neutrophils, Mϕ2 - M2 Macrophages, Mϕ1 - M1 Macrophages, Mϕ0 - M0 Macrophages, EndP - Endothelial Precursor cells, MK - Megakaryocytes, Ery - Erythroblasts, VProm - 'Virtual Promoter' and Coding - variants found in protein coding regions of target gene. After each label is the leaf node specific gene score (lCOGS). The method prioritises a causal mechanism in activated $CD4^+$ T cells for *AHR*.

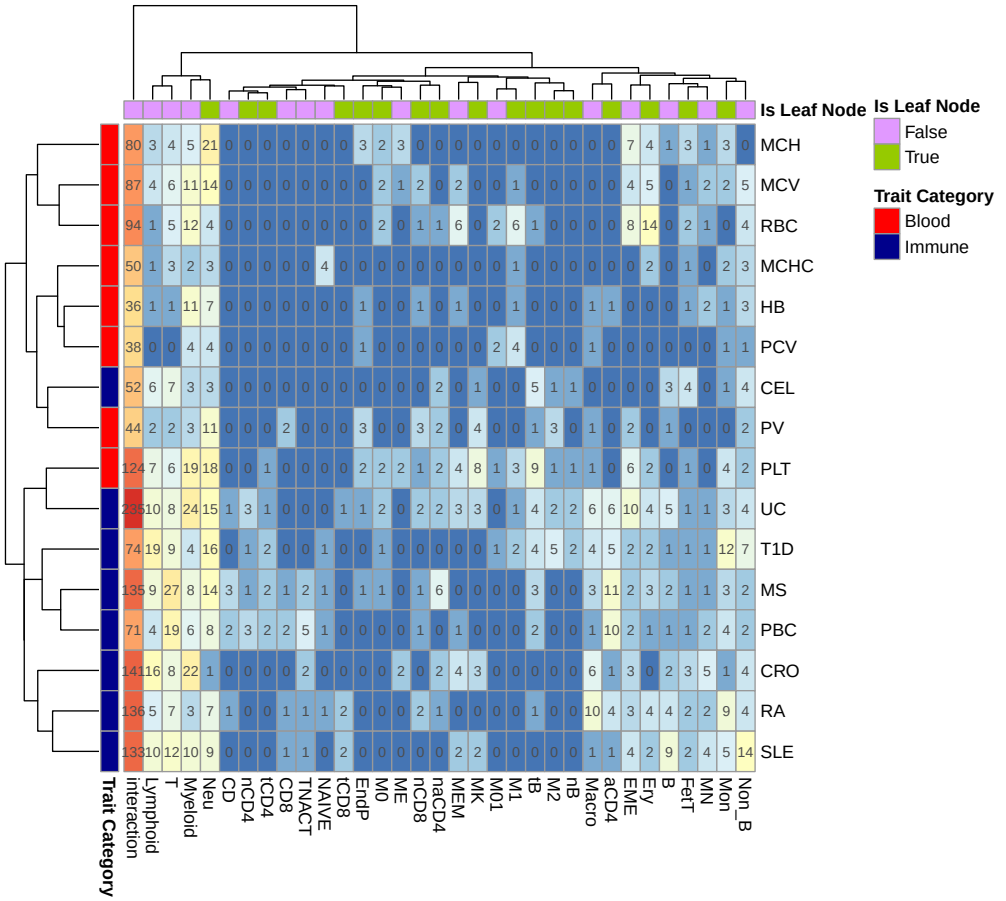


Fig. 3.15 Heatmap of hierarchical COGS analysis using COGS threshold > 0.5. Rows and columns show traits and node labels respectively (See Figure 3.13), with individual cells annotated with protein coding gene counts 'stuck' at a node for a trait, with blue(0 genes) to red (235 genes) visually indicating this information on the $\log_{10}+1$ scale. Additionally rows are annotated with their trait category and columns with their leaf node status (i.e. whether a single or union of tissues). Rows and columns are hierarchically clustered using 'average' linkage.

size, given that genes are prefiltered on overall COGS score. Indeed in the previous chapter *blockshifter* analysis showed that CEL had the second highest enrichment of GWAS signals in activated $CD4^+$ T cells, where as no genes are specifically prioritised in that tissue in this analysis, indicating that COGS score thresholding may be obscuring some of the signal.

Most genes were unable to be mapped to a specific tissue context and instead were prioritised by the non-specific 'interaction' class (mean 95.6). Whilst there was some specificity this was mostly observed within collections of tissue contexts (e.g. 'Myeloid'), rather than at 'leaf' nodes, and overall whilst blood and immune traits do cluster separately, it is challenging to separate them visually. A further

challenge is that a PIR can interact with more than one gene promoter, and as previously discussed a proportion of captured promoter fragments contain more than one gene promoter. This leads to the tissue specific prioritisation of multiple genes, and whether this is due to underlying biology or a limitation of PCHi-C resolution remains to be elucidated. Overall this suggests pinpointing specific cellular contexts using PCHi-C input alone is difficult, supporting the inclusion of additional genomic information prior to functional followup, which I discuss further in Section 3.8.5.

3.8 Cell context specific COGS analysis of immune-mediated disease

Notwithstanding the evidence presented for a role for $CD4^+$ T cell chromatin organisation in immune-mediated disease, differences in DNA methylation of immune-related genes have been observed in $CD4^+$ T cells from autoimmune disease patients compared to healthy controls (Coit et al., 2013; Paul et al., 2016). $CD4^+$ T cells are at the centre of the adaptive immune system and exquisite control of activation is required to guide a $CD4^+$ T cells fate through selection, expansion and differentiation into one of a number of specialised subsets (Murphy and Weaver, 2017). The analysis of gene expression in $CD4^+$ T cells from 20 healthy individuals across a 21 hour activation time course supported early (< 4 hours) modulation of transcriptional programs to be of particular importance, as this is when a majority of differential gene expression is observed (Burren et al., 2017).

These multiple strands of evidence for the importance of early $CD4^+$ T cell activation, coupled with a desire to further understand how disease variants might operate in a particular cell context specific environment, motivated me to take a more focused approach to the integration of autoimmmune and autoinflammatory risk loci with PCHi-C maps. To this end I decided to focus on comparing chromatin structure between resting and activated $CD4^+$ T cell within the context of IMD.

The focus on IMD also provided the opportunity to use an alternative collection of association summary statistics and genotype data afforded by the ImmunoChip platform. This genotyping platform is specifically designed to provide dense genotype coverage of approximately 180 regions with robust demonstration of association with one or more IMDs (Cortes and Brown, 2011). The main benefit of using such a platform is that dense genotyping obviates the need to rely on PMI

or imputation in order to provide sufficient resolution, albeit at the cost of reduced genomic coverage.

3.8.1 ImmunoChip study collection description

I gathered publicly available ImmunoChip summary statistics for four IMD studies from ImmunoBase (Table 3.3), selecting traits on the basis of access to supporting individual level genotyping data. I fine mapped these traits using the sCVPP method. I made one important modification by replacing the definition of approximately LD independent regions (Section 2.4.1) with the 187 non-HLA regions (median size 227Kb with an inter quartile range of between 126Kb and 392Kb) that were densely genotyped on the ImmunoChip. Across all four diseases the mean number of variants with summary statistics was 129,006 (IQR 129,499 and 132,323)

Trait	Cases	Controls	Reference
Autoimmune thyroid disease	2,733	9,364	(Cooper et al., 2012)
Rheumatoid arthritis	11,475	15,870	(Eyre et al., 2012)
Type 1 diabetes	8,000	12,272	(Onengut-Gumuscu et al., 2015)
Celiac disease	12,041	12,228	(Trynka et al., 2011)

Table 3.3 ImmunoChip studies integrated with Javierre et al. (2016) PCHi-C datasets. All summary statistics were downloaded from ImmunoBase.

3.8.2 Allowing for multiple causal variants within a locus

One drawback of the fine mapping technique employed up to this point (sCVPP) is that it assumes that within a given locus there are between zero and one causal variants. When such methods are applied to loci where there are more than one causal variant the credible intervals may not contain any causal variants with a much higher than expected likelihood (Hormozdiari et al., 2015).

Where raw genotyping data for a study is available then forward variable selection methods, such as conditional stepwise regression, can be used to elucidate independent causal variants. Such an approach, first selects a single variant that best explains the variance of a trait with subsequent 'steps' seeking other variants that explain additional trait variance *conditional* on this 'top' variant. Such an approach is routinely applied to GWAS and strong evidence for multiple independent causal variants within a locus has been described (Haiman et al., 2007). Whilst

this approach is attractive computationally, doubts exist over its validity (Miller, 1984). This is because such a method is not equivalent to explaining which variants *jointly* explain the variance of a trait. However, approaches that search such a potentially large model space exhaustively are only computationally feasible for simple models incorporating a limited number of variants (Bottolo and Richardson, 2010; Lee et al., 2018) .

An alternative approach is to use Monte Carlo methods to sample the model space allowing the consideration of multiple causal variants within a genetic region. One example is GUESSFM that uses a Bayesian evolutionary stochastic search algorithm to effectively sample the model space, and has been shown to have consistently better performance than step-wise conditional regression approaches when variants are correlated due to LD (Wallace et al., 2015).

3.8.3 Comparison of COGS scores between single and multiple causal variant approaches

I obtained, from Chris Wallace, a previous analysis of four ImmunoChip studies (Table 3.3) using GUESSFM (Wallace et al., 2015). This afforded the opportunity to gauge the effect of the single causal variant assumption on COGS gene scores by comparing the genes prioritised by COGS using GUESSFM and sCVPP computed from summary statistics, that are the usual COGS input.

I limited analysis to PCHi-C contact maps for activated and non-activated $CD4^+$ T cells for reasons detailed in Section 3.8. COGS requires modification to deal with GUESSFM output that consists of vector of posterior probabilities for causal models within a given region, r, such that $\mathbf{p} = (p_1, p_2, \ldots, p_{1-q}, p_q)$ over q models, indexed by i, that themselves can incorporate 1 or more variants. Overall $\sum_{i=1}^{q} p_i \leq 1$ and in this scheme it is important not to double count SNPs, which occurs if COGS is naively applied to marginal posterior probabilities of SNP inclusion, as the same SNP can be present in multiple models. Thus, it is no longer feasible to precompute 'LD aware' fragment posterior probabilities as described in Section 3.7.1, but instead I construct indices of variants genotyped for a given trait and their overlap with sets of features. For example the union of *Hin*dIII fragments incorporating PIRs for activated $CD4^+$ T cells and VPF for a given gene would entail the union of two lookups in order to identify genotyped variants. The posterior probability within a region r for this set of features, f_r, to contain one or more variants, is the sum of the posterior probabilities for any

model containing at least one of these variants

$$\sum_{i:i\in f_r} p_i, \tag{3.3}$$

where $i \in f_r$ indicates that the i^{th} model contains one or more variants that physically overlaps with the the feature set f_r, in this case CD4$^+$ T cell PIRs and VPF. Evidence across m regions can be combined for a gene, g, using

$$\text{mCOGS}_g = 1 - \prod_{j=1}^{m} \left(1 - \sum_{i:i\in f_{r_m}} p_i\right). \tag{3.4}$$

Importantly, this score is comparable to COGS scores computed using the sCVPP method employed in single causal variant COGS analysis.

Disease	sCOGS	mCOGS	Both	Total
Type 1 diabetes	16	19	33	68
Coeliac disease	21	2	35	58
Rheumatoid arthritis	16	7	17	40
Autoimmune thyroid disease	8	4	6	18
Total (%)	61 (33%)	32 (17%)	91 (50%)	184

Table 3.4 Protein coding gene count comparison between COGS prioritised genes (COGS score > 0.5) with different inputs. sCOGS - count of genes prioritised exclusively using as input sCVPP, mCOGS - count of genes prioritised exclusively using marginal posterior probabilities for a set of variants to be causal from GUESSFM analysis of raw genotyping data. Both - count of genes prioritised using both input methods.

I ran parallel COGS analysis with two inputs, sCVPP and GUESFM, across the four diseases, calling these sCOGS and mCOGS scores respectively. I restricted analysis to ImmunoChip regions where data were available for both inputs. These analyses generated an 'overall' COGS score that incorporated evidence from PCHi-C data contact maps for activated and non-activated CD4$^+$ T cells along with coding SNP and VPFs. Across all four traits this prioritised (COGS score > 0.5) 152 and 123 protein coding genes for sCOGS and mCOGS respectively (Table 3.4). Overall there was modest agreement between the two input methods with 91 genes prioritised by both methods.

As expected when genes are prioritised by both input methods, COGS scores are correlated, and overall, across all diseases combined, Pearson's correlation coefficient is 0.35 (Figure 3.16). There are, however, clusters of genes prioritised by one method, where the COGS score in the alternative method is close to zero.

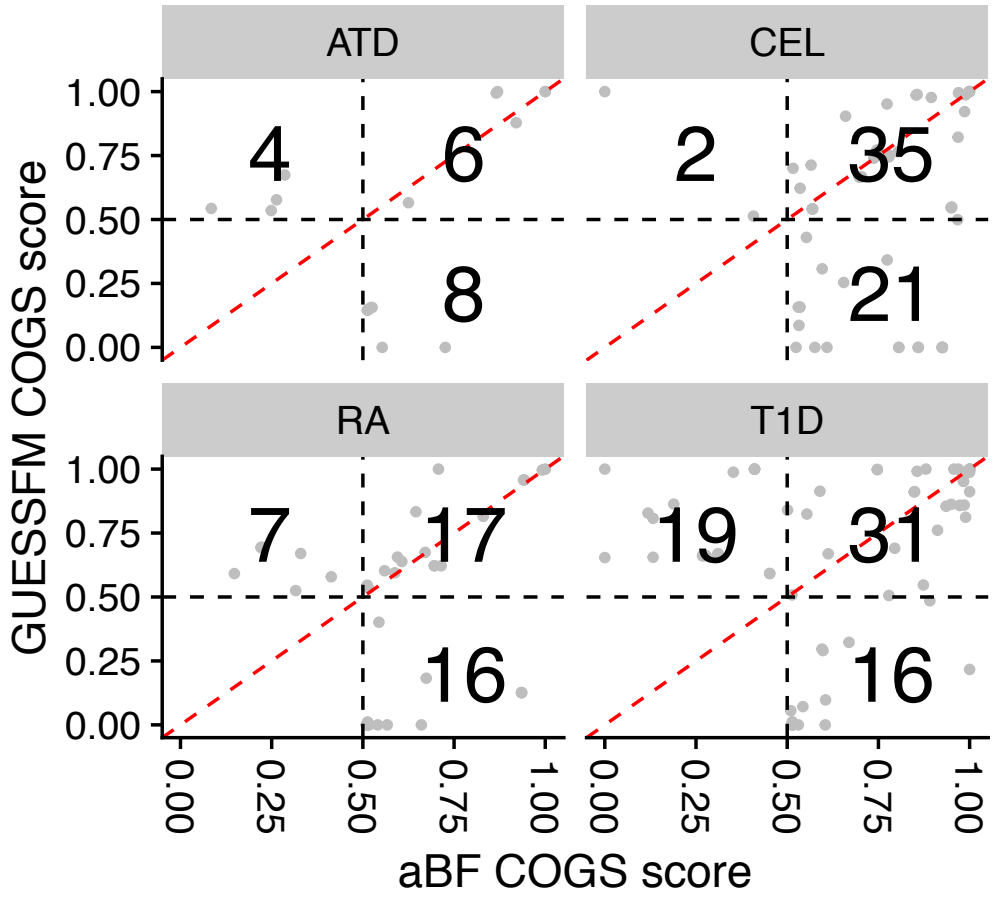


Fig. 3.16 Comparison of 'Overall' COGS scores for sCVPP and mCVPP (using GUESSFM) fine mapping inputs. ATD - Autoimmune thryoid disease (Cooper et al., 2012), CEL - Coeliac disease (Trynka et al., 2011), RA - Rheumatoid arthritis (Eyre et al., 2012) and T1D - Type 1 diabetes (Onengut-Gumuscu et al., 2015). Only genes prioritised (COGS score > 0.5) by at least one method are shown. Numbers reflect the counts in each of the relevant quadrants.

One of the aims of COGS is to facilitate functional follow up by ranking genes at a given locus. For each disease and region pair, I ranked genes using either sCOGS or mCOGS scores. Out of a total of 79 disease-region pairings examined, there was an 80% agreement in the highest scoring gene.

I conducted a more detailed analysis of an exemplar region, 19p13.2, in T1D, for which multiple genes were prioritised using sCOGS input but not mCOGS (Figure 3.17). In this region strong functional evidence for the causal candidacy of *TYK2* has been reported (Dendrou et al., 2016) and multiple independent signals within the locus have been previously described (Onengut-Gumuscu et al., 2015). Both input methods give the highest priority to *TYK2* selected rs34536443

as a likely causal variant in agreement with Onengut-Gumuscu et al. (2015). Additionally, the sCVPP input prioritises a putative causal variant (rs144309607) in the virtual promoter region of *TYK2*, which itself is imputed using the PMI method. This single alternative signal is sufficient to prioritise five additional genes through PCHi-C interactions explaining the discrepancy and it is likely that this sCOGS specific signal is an artefact of the PMI method. The mCOGS approach also benefits from allowing multiple causal variants: GUESSFM prioritising the independent coding (missense) signal (rs12720356) in agreement with Onengut-Gumuscu et al. (2015), amplifying the overall COGS score for *TYK2*.

The T1D susceptibility region at 16p11.2 shows a reciprocal pattern; mCOGS prioritises five genes which sCOGS fails to prioritise at a COGS score > 0.5 (Figure 3.18). One explanation for this is that the assumption of a single causal variant, results in a diffuse signal spread across many variants resulting in lower overall COGS scores. In contrast, GUESSFM picks two independent signals (rs151233 and rs151234) amplifying gene scores accordingly.

3.8.4 Cataloguing PCHi-C prioritised genes across immune-mediated disease

Due to the modest overlap between mCOGS and sCOGS inputs and in the absence of compelling evidence for the superiority of one input method I ran hierarchical COGS over the ImmunoChip datasets using both inputs for PCHi-C maps of activated and non-activated $CD4^+$ T cells, taking forward the union of prioritised genes. To augment this I added GWAS data for SLE (Bentham et al., 2015) and RA (Okada et al., 2014) for which imputed summary statistics were available in order to create a 'long' list of immune-mediated disease prioritised genes using the sCVPP finemapping input method. In total I prioritised 245 unique protein-coding genes across all datasets and analyses (Figure 3.19).

3.8.5 Integration of functional data with COGS scores

Having generated a long list of genes prioritised from the integration of GWAS, ImmunoChip and PCHi-C datasets, I sought to understand whether orthogonal functional genomic data might be used for to refine putative causal mechanisms. I obtained enhancer RNA (eRNA) annotations, and total RNA sequencing data, collected and analysed by Tony Cutler and Arcadio Rubio-Garcia from the same donors for which PCHi-C was performed, for both activated and non-activated

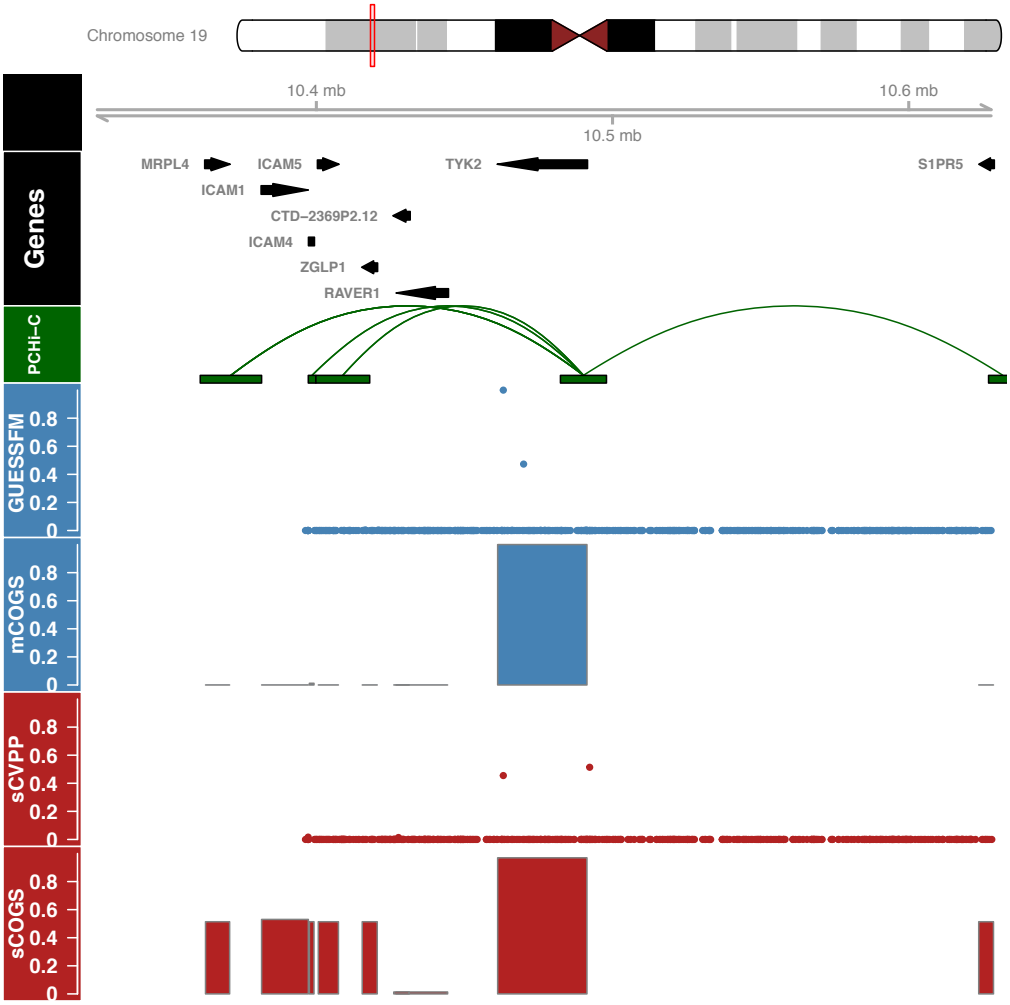


Fig. 3.17 Comparison of T1D (Onengut-Gumuscu et al., 2015) COGS scores derived from sCVPP and GUESSFM inputs at the 19p13.2 susceptibility locus. For clarity only protein coding genes with a COGS score > 0.01 from either input are shown. The dark green stanza shows the union of interactions with the *TYK2* promoter in activated and non-activated $CD4^+$ T cells, for clarity other interactions in the region are not shown. The blue 'GUESSFM' stanza shows the marginal posterior probabilities for SNPs to be causal allowing for multiple causal SNPs. The blue 'mCOGS' stanza presents the overall COGS scores computed from GUESSFM model posterior probabilities for a given gene (see 'Genes' stanza). Similarly the red stanzas present sCVPP and resultant sCOGS scores from allowing a maximum of one causal variant.

$CD4^+$ T cells. Briefly, eRNAs are defined as bi-directionally transcribed RNA species that map to ChIP-Seq derived enhancer regions but do not map to known gene annotations (Li et al., 2016). This bi-directionality is key to the robust identification of intronic eRNAs as it allows their separation from pre-mRNA on

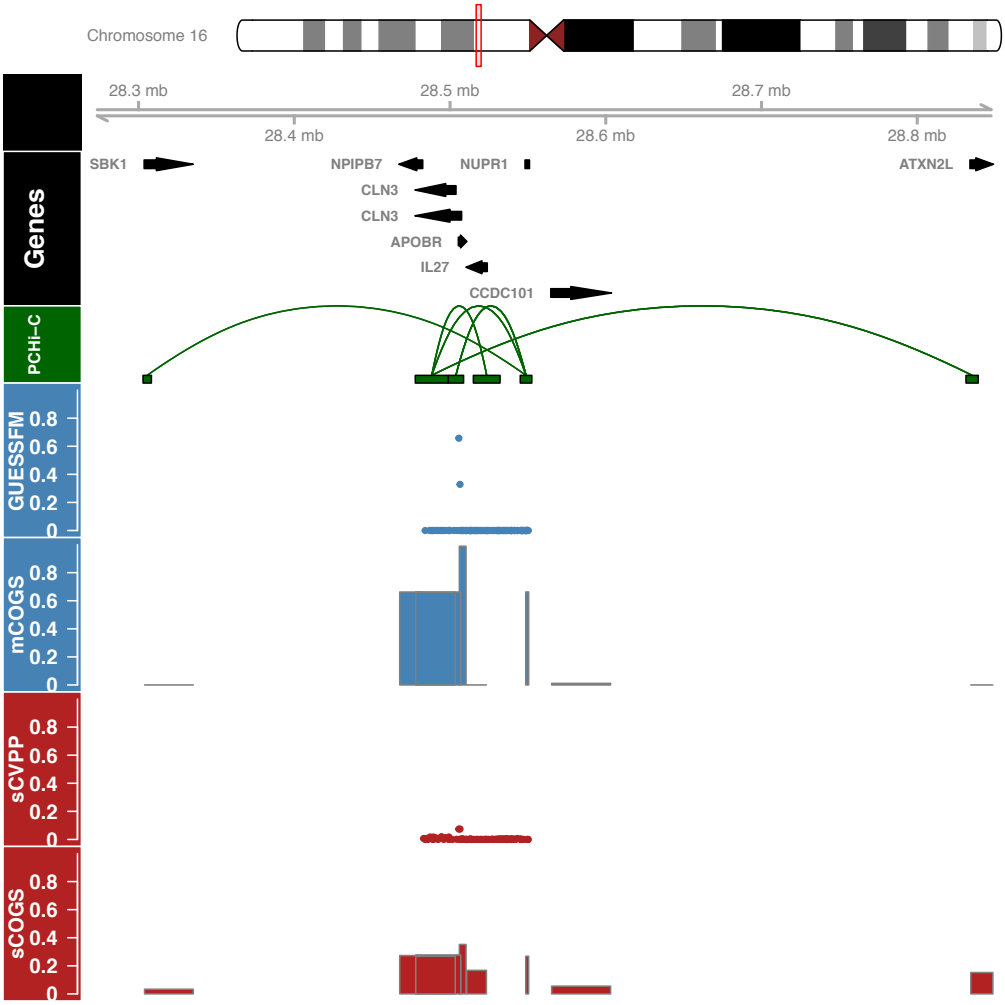


Fig. 3.18 Comparison of type 1 diabetes (Onengut-Gumuscu et al., 2015) COGS scores derived from sCVPP and GUESSFM inputs at the 16p11.2 susceptibility locus. Stanzas are described in the legend of Figure 3.17. All activated and non-activated CD4$^+$ T cells interactions for genes with COGS > 0.5 are shown.

the basis of strand. I applied a stepwise filtration based on the integration of this data.

Of the 245 putative causal genes prioritised in the previous section (Section 3.8.4) 179 were expressed in at least one CD4$^+$ T cell activation state on the basis of RNA-seq data. Of these, 118 were proximal to a GWAS significant index variant ($p < 5 \times 10^{-8}$) through a PCHi-C connection. Within this set of 118 genes, 63 (48%) lay outside of the disease susceptibility region by which they were prioritised. Examples include, *IL6ST* (124 kb from 5q11.2 susceptibility region) in RA (Stahl et al., 2010) and *GPR183* (76kb from 13q32.3 susceptibility region)

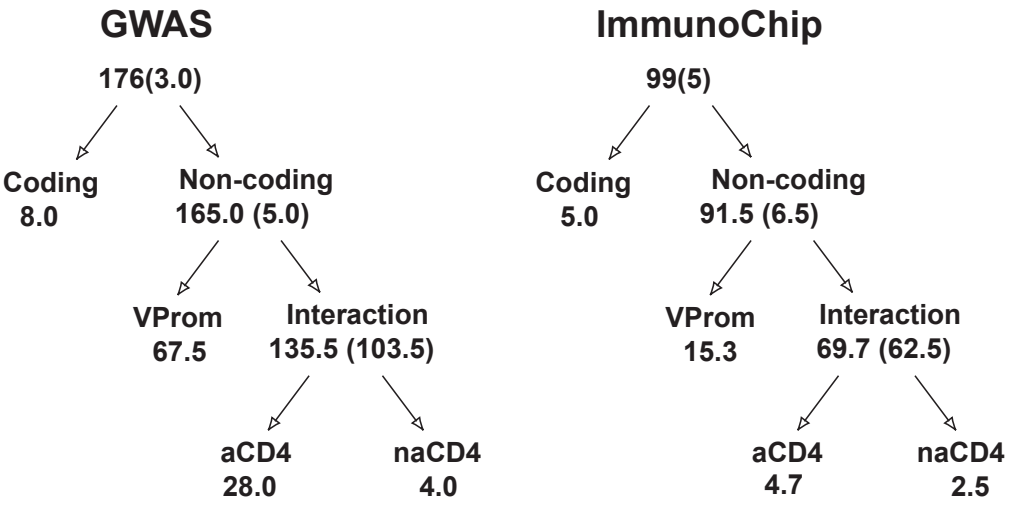


Fig. 3.19 Results of running hierarchical COGS across five immune-mediated diseases. On the left are results using input from imputed GWAS datasets (Bentham et al., 2015; Okada et al., 2014) using sCVPP, on the right are the results using ImmunoChip datasets (Table 3.3). Labels indicate the prioritised causal mechanism. When the same gene was prioritised for multiple diseases, I assigned fractional counts to each node, defined as the proportion of the n diseases for which the gene was prioritised at that node. Numbers in brackets indicate the fractional count of genes assigned to a non leaf node.

in T1D (Heinig et al., 2010; Wallace et al., 2012). Overall, the mean distance from peak GWAS signal to a prioritised gene was 153kb.

Of these 118 genes, 82 were differentially expressed between activation states (FDR < 0.01) and 48 were prioritised due to an interaction identified only in aCD4 or nCD4. Sixty-three genes were connected via 'PCHi-C' to a fine-mapped variant that overlapped an expressed eRNA (Burren et al., 2017).

3.8.6 Functional validation in *IL2RA*

One of the genes priortised in multiple immune-mediated diseases that also appeared in all the functional categories previously described (Section 3.8.5) was *IL2RA* (Eyre et al., 2012; IMSGC et al., 2013; Jostins et al., 2012; Onengut-Gumuscu et al., 2015). *IL2RA* encodes the CD25 protein, a component of the IL-2 receptor that is essential for high-affinity binding of IL-2, regulatory T cell survival and T effector cell differentiation and function (Liao et al., 2013). I found this prioritisation to be driven by an interaction between the *IL2RA* promoter and a PIR in intron 1 known to harbour a set of type 1 diabetes putative causal variants (red group in Figure 3.20) identified in a previous fine mapping study (Wallace

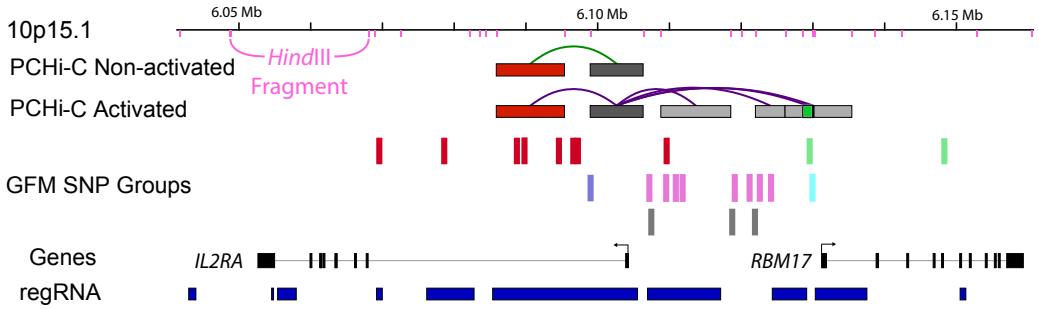


Fig. 3.20 Genomic and genetic architecture of 10p15.1 type 1 diabetes susceptibility locus. PCHi-C interactions link the *IL2RA* promoter to autoimmune disease associated genetic variation which leads to expression differences in *IL2RA* mRNA. GUESSFM SNP groups ('GFM SNP Groups' stanza) from (Wallace et al., 2015) are partitioned by colour. PCHi-C data is for activated (green) and non-activated (purple) $CD4^+$ T cells, in unambiguous cases PIRs are coloured by GUESSFM SNP group overlap. Figure adapted from an original prepared by Tony Cutler, Arcadio Rubio Garcia and Chris Wallace.

et al., 2015). This set of variants (Figure 3.20) is in high LD ($r^2 > 0.8$) with rs12722495 which has been shown to associate with the surface expression CD25 in memory T cells (Dendrou et al., 2009).

Using a targeted RNA-sequencing approach, and software I helped to develop previously (Rainbow et al., 2015), Daniel Rainbow measured the relative expression of *IL2RA* mRNA in five individuals heterozygous across the red group of SNPs (Figure 3.20 who were also homozygous across most other associated SNPs, in a 4 hour activation time course of $CD4^+$ T cells. Allelic imbalance was observed consistently for two reporter SNPs in intron 1 and the $3'$ UTR in non-activated $CD4^+$ T cells in each individual (Figure 3.21a) validating a functional effect of the PCHi-C derived interaction between this PIR and the *IL2RA* promoter in non-activated $CD4^+$ T cells. While the allelic imbalance was maintained in non-activated cells cultured for 2-4 hours, the imbalance was lost in cells activated under *in vitro* conditions. Since increased CD25 expression with rare alleles in the red group of SNPs has previously been observed on memory $CD4^+$ T cells but not naive or T regulatory subsets that are also present in total $CD4^+$ T cell population, we purified memory cells from eight red group heterozygous individuals and confirmed activation induced allelic imbalance of *IL2RA* mRNA expression in this more homogeneous population (Figure 3.21b). One possible explanation for this is that formation of additional chromatin loops with alternative GUESSFM SNP groups (blue, magenta, grey, cyan and green Figure 3.20) on activation overcomes the basal effect of the more constitutive enhancer effect modulated by contact

within intron 1. Further support for rs61839660 to be causal has subsequently been reported (Huang et al., 2017; Rainbow et al., 2017). This empirical analysis confirms that the identified PIR contains a variant which functionally affects the transcription of the identified target gene (Burren et al., 2017).

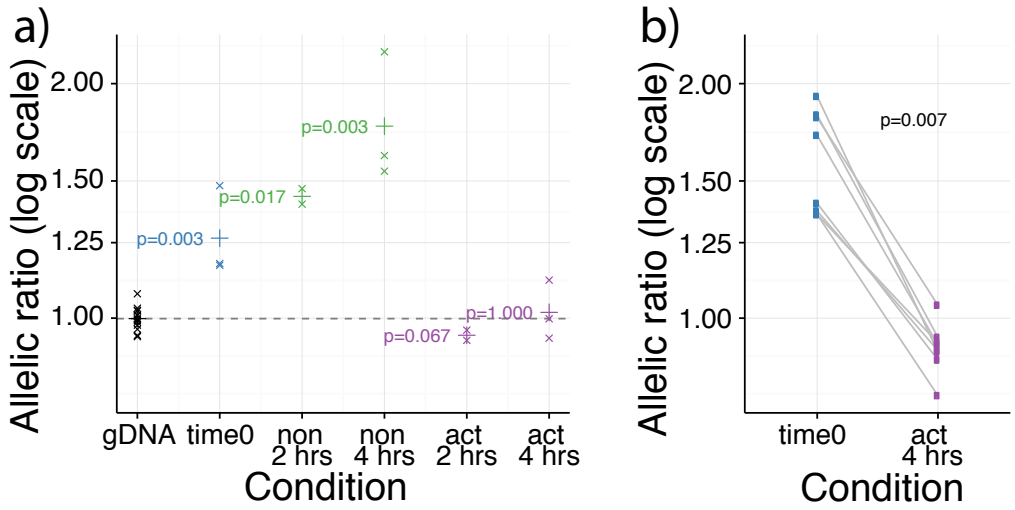


Fig. 3.21 **a)** Allelic imbalance in mRNA expression in total CD4$^+$ T cells from individuals heterozygous for the red (Figure 3.20) group of GUESSFM SNPs using rs12722495 as a reporter SNP in non-activated (non) and activated (act) CD4$^+$ T cells cultured for 2 or 4 h, compared to genomic DNA (gDNA, expected ratio = 1). Allelic ratio is defined as the ratio of counts of T to C alleles. 'x' represents geometric mean of the allelic ratio over 2–3 replicates within each of 4–5 individuals; *p*-values from a Wilcoxon rank sum test comparing complementary DNA (cDNA) to gDNA are shown. '+' shows the geometric mean allelic ratio over all individuals. **b)** Allelic imbalance in mRNA expression in memory CD4$^+$ T cells differs between ex vivo (time 0) and 4-h activated samples from eight individuals heterozygous for red (Figure 3.20) group of GUESSFM SNPs SNPs using rs12722495 as a reporter SNP. *p*-value from a paired Wilcoxon signed rank test is shown

3.9 Discussion

Although GWAS have been successful in discovering a multitude of trait associated loci, underlying causal mechanisms have proved to be more elusive. Whilst, LD impedes the identification of causal variants, so their regulatory nature often complicates functional assessment of causal genes. Often to fill this knowledge vacuum, researchers resort to more *ad hoc* rather than data-driven methods for suggesting causal candidate genes and the tissues in which they might operate.

The results presented in this chapter suggest that PCHi-C gene prioritisation using the COGS method developed perform better than the proximity based methods that have been widely adopted in the field (Section 3.6.1). This increased performance of PCHi-C COGS was evident, in the smaller total number of genes prioritised, and the suggestion of causal genes that were not found using either proximity or TAD based inputs (Section 3.6.2).

I found that by adapting COGS to generate prioritisation scores in a tissue dependent manner I was able to identify, albeit for only a modest number of cases, evidence for the importance of specific tissue context and trait combinations (Section 3.7). Overall hierarchical clustering of these results was inconclusive using the canonical COGS threshold (COGS>0.5) although I obtained more robust results that discriminated blood and immune traits when this was lowered (COGS>0.01). Leaving aside technical considerations, this apparent lack of tissue specificity might be explained by the fact that the transcriptional programs by which causal variants modulate disease risk might be active across a wide range of contexts and therefore difficult to resolve at the individual tissue level.

Of course there are a wide range of technical reasons for these results which warrant discussion. In order to explore as many traits as possible, I employed an LD based method, PMI, in order to increase SNP coverage across GWAS datasets. Using sCVPP derived from PMI as COGS input is likely to lead to a less robust gene prioritisation for a number of reasons. Consider the instance where a causal SNP is not directly typed and thus no association summary statistics are directly available, in such a case the association can only be measured by 'proxy' variants in LD. Generally, as LD between this 'proxy' SNP and the causal SNP decays so too will the association. In this example, PMI will assign the association p-value of the 'proxy' to the untyped causal variant, attenuating the signal, reducing downstream fine-mapping resolution, ultimately leading to a degradation in COGS prioritisation. The opposite is also true, if the causal variant is typed, as untyped variants in LD, by PMI, will be assigned the same p-value. In regions of high LD this will exacerbate the phenomenon of LD shelving, that occurs when many SNPs are in high LD with the causal variant, such that they all have similar association summary statistics (illustrated in Figure 3.17). Whilst such LD shelving is a true biological phenomenon that is recognised to limit all fine-mapping methods, PMI amplifies this, resulting in an increased uncertainty in gene prioritisation, that manifests as a larger number of genes which themselves have lower prioritisation scores.

I found the assessment of which input method, sCOGS or mCOGS performed best challenging due to the lack of a gold standard of confirmed causal genes underlying disease susceptibility. One might expect that differences in performance between the two input methods to be somewhat dependent on how often the single causal variant assumption is violated. Recently, Asimit et al. (2019) used a Bayesian multinomial stochastic search method (MFM), to simultaneously fine-map autoimmune disease susceptibility loci across six autoimmune diseases. They found that their fine-mapping results within regions with the strongest biological prior (i.e. containing known genes that regulate T-cell function) and effect were more likely to be discordant with the more widely applied stepwise conditional regression approach widely employed. One interpretation of this is that the phenomena of multiple independent variants within a disease associated regions is likely to exist genome-wide, but is currently hidden due to insufficient study power or sample size. With this in mind I would expect sCVPP input to perform less well than the GUESSFM input for those regions and overall and this might explain the more modest overlap between the sCVPP and GUESSFM input methods that I observed.

Underscoring the trade off, when integrating genomic and GWAS data, are the logistics of data access; sCVPP requires only summary statistics and thus can integrate a large selection of traits, whilst GUESSFM requires individual level data limiting its application. It seems likely that using the fine-mapping results of Asimit et al. (2019) as COGS input might provide a more robust set of prioritised causal genes than the input methods I have presented although further work is required to investigate this.

As well as the limitations due to fine-mapping input, a discussion of PCHi-C limitations and how these might effect causal gene prioritisation is required. The most serious, involves the blind spot for observing shorter range interactions that involve *Hin*dIII fragments and adjoining baited interactions and I attempted to capture these as virtual promoter fragments (VPF). This approach whilst reasonable might result in the inappropriate assignment of variants to a gene given the expected size of a VPF is approximately 27kb. From imputed GWAS and iChip COGS prioritisation I found that 38% and 15% of genes were prioritised by VPF respectively (Figure 3.19), a sizeable fraction of the total number of unique genes selected. Whilst incorrect assignment of variants to a gene on the basis of VPFs will result in more genes prioritised adding noise to the gene list, it would not by itself disadvantage true causal genes from being prioritised. In such cases the integration of additional sources of genomic information, for example chromatin

accessibility and gene expression data prior to selecting genes and tissue contexts for functional followup is essential (Kumasaka et al., 2019; Pickrell, 2014).

A second more pernicious issue is that many baited fragments are promiscuous in that they contain promoter regions for more than one gene. For these promiscuous baits it is impossible to resolve which promoter or promoters are involved in the chromatin looping, using the current PCHi-C data alone. One approach is to use gene expression patterns to provide a filter for target genes identified by specific chromatin interactions, which I discuss in section 3.8.5. Furthermore, my analysis to date has concentrated on protein coding genes as these have the most mature and complete annotation, however, research has suggested a role for non coding genes in the modulation of autoimmune disease susceptibility (Castellanos-Rubio et al., 2016). The PCHi-C platform used does provide some coverage of the non protein coding genome, however this is not exhaustive by design and due to the overlapping nature of the coding and non-coding genome, the issue of promiscuous baits is worsened.

The introduction of an alternative restriction enzyme that cuts at greater frequency thus resulting in, on average, shorter fragments, might overcome some of these issues (Chesi et al., 2018). It will however introduce additional challenges; the shorter fragments will make it more difficult to identify unique sequence for capture design, exacerbated by the increased number of fragments to be captured, which in turn will require more sequencing in order to provide adequate sequence coverage.

Another limitation is the threshold approach to CHiCAGO scores that are used to call interactions. All of the methods developed so far use a threshold score of 5, so that an interaction with a score of 4.99 will be omitted. I am aware of an alternative method, based on a Bayesian sparse variable selection approach (Eijsbouts et al., 2019), that will allow the assignment of posterior probabilities to interactions that might obviate the need for this threshold approach. An extension of COGS would be required in order to integrate the two posterior distributions (location of causal variants and PCHi-C contacts) and this is a promising avenue for future research.

Excluding single cell implementations, all genomic technologies give an average of the molecular events across the (sometimes mixed) population of cells being assayed. In the case of immune subsets this is particularly relevant as broad categories, such as $CD4^+$ T cells will be heterogeneous containing further subdivisions that may or may not be relevant for disease biology. It is important to bear this in mind when considering PCHi-C maps, as without single cell profiling (Tan

et al., 2018), it is impossible to resolve whether interactions are common across the assayed tissue type or are specific to, an underlying, and as yet unsorted, subset.

Given the limitations of PCHi-C and the differences between input methods and thus the veracity of the genes prioritisised, I think it important that further genomic information is taken into context before embarking on further functional studies. Ideally such genomic data should be for the same cellular contexts and if possible the same individual donors for which PCHi-C was performed. In my analysis the integration of chromatin state and RNA expression data enabled further filtering allowing, a reduction from 245 putative candidate genes, identified from combining results using both sCVPP and GUESFM COGS results across immune-mediated diseases, to 118, found to be expressed in either activated or non-activated $CD4^+$ T cells (Section 3.8.5). I was able to reduce this further by considering specific annotations such eRNA overlap and context specific interactions. Not only did such integration reduce the number of genes but it also has the benefit of suggesting causal mechanisms amenable to followup functional experiments, and I described one such example involving the gene *IL2RA*. These results suggest an overarching framework to GWAS causal candidate gene prioritisation, that incorporates an initial scan with *blockshifter* to identify target tissues of particular interest followed by detailed finemapping and subsequent integration, using COGS with relevant tissue types.

The example of *IL2RA* demonstrates how complicated and subtle the mechanisms by which common variants modulate disease risk. In this associated region, multiple independent variants that in some cases can be shared and at other times distinct, between different immune-mediated diseases act to affect the expression of *IL2RA* and potentially other causal genes in this region (e.g. *RBM17*). Results in this region show the importance of tissue context in disease mechanisms, as $CD4^+$ T-cell activation was sufficient to overcome the allelic affect observed for rs61839660 in non-activated cells. This demonstrates the challenges that are involved in elucidating causal mechanisms in common disease, in that even for a data set of 17 cell types with known relevance to IMD, only a small minority of cellular contexts have been assessed. It is therefore almost certain that we are missing key contexts in which many common variants act in order to modulate disease risk. Whilst the framework set out here will be useful, I emphasise that the integration of multiple strands of evidence and empirical techniques are necessary for the robust identification of causal genes and tissue contexts and ultimately underlying disease mechanisms.

My analysis adds to a growing body of evidence that regulatory variation associated with complex disease is often subtle and context specific. To date, there has been a focus on methods that integrate GWAS and eQTL data as discussed in section 3.2.2. Due to logistical, technical and economic barriers, the generation of well powered catalogues of eQTLs across a range of tissue contexts has proved challenging. Non-population based techniques, such as PCHi-C, that do not require a large amount of resource, have utility as they can be used to probe a larger set of disease-relevant tissues and contexts. Indeed, subsequent to the work described here many studies have been completed in other disease relevant tissues including; pancreatic islets cells for type 2 diabetes (Miguel-Escalada et al., 2018), cardiomyocytes for heart disease (Choy et al., 2018), iPSC osteoblasts for bone mineral density disorders (Chesi et al., 2018) and iPSC-induced neurons for neurological disorders (Song et al., 2018). Another promising application of COGS, that I am developing, is to use pleiotropy between common and rare mono/oligo-genic disease to suggest novel causal genes in the latter, for example between IMDs and primary immune deficiency (Thaventhiran et al., 2018).

In summary methods, such as COGS, that link putative causal variation with target genes, show great promise in prioritising disease mechanisms and tissue contexts. Along with orthogonal population based methods they provide a powerful and data driven approach to informing the design of, but not replacing, detailed empirical approaches for the characterisation of the causal molecular events underlying human disease.

Chapter 4

Shared and distinct genetic architectures in immune-mediated disease

4.1 Foreword

4.1.1 Chapter Summary

In this chapter I develop a principal component analysis based framework that uses GWAS summary statistics to summarise the similarities and differences in genetic architecture between a set of clinically related diseases. I apply this framework to a set of immune-mediated diseases (IMDs) in order to generate a lower-dimensional 'basis' that summarises their combined genetic architectures into a set of ordered principal components (PC). I attempt to characterise PCs by projecting on summary statistics for a wide range of binary and quantitative traits. Next, I apply the method to a GWAS of juvenile idiopathic arthritis (JIA) disease sub-types where cohorts sizes are small, in order to characterise their shared and distinct genetic architectures. Finally I turn my attention to updating the proposed framework to allow the projection of individual level genotype data onto the basis. With this I attempt further characterisation of basis components using eQTL datasets targeting relevant primary immune cell types.

4.1.2 Attributions

The work presented in this chapter relies on an unpublished case/control cohort examining the genetics of JIA disease subtypes. Genotype and phenotype data

was supplied by John Bowes, Sam Smith, Annie Yarwood, Damian Tarasek and Wendy Thomson and I had no part in study design, recruitment or genotype quality control.

All UK BioBank GWAS summary statistics were generated by the Neale Laboratory and were downloaded from http://www.nealelab.is/uk-biobank/.

4.1.3 Motivation

After initial concerns about privacy (Homer et al., 2008), summary statistics for well powered GWAS are becoming increasingly available in the public domain (MacArthur et al., 2017). For immune-mediated diseases that cluster in individuals or families, there is strong evidence for a shared genetic architecture (Cho and Feldman, 2015; Cotsapas et al., 2011; Onengut-Gumuscu et al., 2015; Smyth et al., 2008; Zhernakova et al., 2009). Whilst studies and methods have been developed to exploit this in order identify pleiotropic loci across multiple diseases (Solovieff et al., 2013), there have been few attempts to integrate summary GWAS data across multiple phenotypes in more holistic fashion at the genome-wide scale.

The lack of methods in this areas coupled with the increasing availability of GWAS summary statistics motivates the development of novel approaches that are able to generate useful summaries of input traits, that in turn can be used for disease classification and characterisation (Cortes et al., 2017). Furthermore, such summaries learned from large well powered GWAS, might facilitate the characterisation of the genetic architecture of non input traits. This is particularly attractive in rare or clinically heterogeneous immune-mediated disease where cohort sizes are typically modest and conventional GWAS approaches are under powered, as such an approach has the potential to provide insights into the shared and distinct genetic architectures which might impact therapeutic intervention.

4.1.4 Software Availability

In the course of this work I developed ‘cupcake’, an R library of functions that encapsulates the methods described in the chapter. This package, is freely available from https://github.com/ollyburren/cupcake under an MIT license agreement.

4.2 Background

Immune mediated diseases, whilst diverse, have been found to co-occur both within individuals and families supporting the presence of a shared genetic architecture (Somers et al., 2006). For example, between 4-9% of individuals with type 1 diabetes (T1D) are also affected by coeliac disease (CEL) (Cronin et al., 1997). Whilst the studies supporting these claims are affected by ascertainment bias, more recent evidence from GWAS has provided a stronger foundation for such assertions (Li et al., 2015; Parkes et al., 2013; Zhernakova et al., 2009). This has lead to concerted efforts to understand both the similarities and differences between the genetic architectures of IMDs, which can be divided into three main approaches which I discuss in more detail below.

Cross-phenotype analysis and colocalisation

Soon after the publication of the first large scale multi-trait GWAS (Wellcome Trust Case Control Consortium, 2007) Smyth et al. (2008) presented a systematic pairwise analysis between T1D and CEL susceptibility loci, finding convincing evidence that seven non-HLA loci were common between traits. Two of these seven loci showed evidence of heterogeneity in effect direction between the two diseases indicating that the same locus could be both protective and modulate risk in different disease contexts. However, at this early stage of GWAS development genotype platform density was low, which precluded inference as to whether shared loci were a result of shared or distinct causal variants.

Cotsapas et al. (2011) performed a detailed study of seven separate IMDs (CEL, Crohn's disease (CRO), multiple sclerosis (MS), psoriasis (PSO), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE) and T1D) selecting SNPs or their proxies that were genome-wide significant in one or more disease studies. They developed the Cross-Phenotype Meta Analysis (CPMA) statistic; using summary GWAS p-values in order to detect evidence for association of a SNP to multiple disease phenotypes. On applying CPMA to 107 SNPs, selected due to genome-wide significance in one or more of the underlying studies, they found evidence for sharing, between two or more diseases at 42 SNPs. Furthermore they described 9 SNPs for which they found strong evidence for opposing effects between diseases at the same locus, indicating that such a phenomenon might be widespread across IMDs.

Whilst both Smyth et al. (2008) and Cotsapas et al. (2011), present compelling evidence for pervasive shared susceptibility loci across IMD, they are limited to testing each SNP independently and are unable to make inferences about whether sharing is due to the same or different causal variants.

An alternative approach presented in (Fortune et al., 2015), that overcomes this is to take a more targeted approach and examine evidence for colocalisation (Giambartolomei et al., 2014) at associated loci between pairs of traits. In this study raw genotype data for four autoimmune traits (RA,T1D,CEL and MS), that had been typed on the ImmunoChip platform was assembled and assessed pairwise for colocalisation. Out of 90 regions assessed, significant evidence supporting colocalisation was found at 33. In a related approach Pickrell et al. (2016) used summary statistics to investigate, genome-wide, the evidence for shared causal variants across 42 GWAS for a broad range of phenotypes and diseases. They identified over 300 loci where there was evidence for a shared locus.

Genetic correlation

An alternative approach to quantify the genetic overlap is to enumerate the genetic correlation between two traits genome-wide. This requires the estimation, for each trait, of the SNP level heritability, that is the proportion of the variance of the trait explained by a given SNP (Section 1.3.1).

If raw genotyping data is available and sample sizes are not prohibitively large the most accurate method of estimating SNP heritability is to apply a linear mixed model, as implemented in the GCTA software package (Yang et al., 2011). More recently LD score regression (Bulik-Sullivan et al., 2015), a method which requires only summary GWAS statistics and can be applied efficiently to very large GWAS has been developed. At the foundation of LD score regression is the assumption that causal variant effect sizes are inflated due to LD with each other. The LD score for a variant j, is defined as

$$\ell_j = \sum_k r_{jk}^2, \tag{4.1}$$

where r^2 is the correlation between variant j and the k^{th} variant. Under a polygenic model the expected value of $z_{1j}z_{2j}$ for the z scores for two traits denoted by subscript 1 and 2 at variant j is

$$\mathbb{E}\left[z_{1j}z_{2j}\ell_j\right] = \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_j + \frac{\rho N_s}{\sqrt{N_1N_2}}, \tag{4.2}$$

where N_1 and N_2 are the sample sizes of the traits, N_s is the number of samples shared between both studies, ρ_g and ρ are, respectively, the genetic and phenotypic correlation between the traits and M are the total number of SNPs being considered. This means that ρ_g can be estimated by the slope of the regression of $z_{1j}z_{2j}$ on LD score ℓ_j. This approach has been used to analyse the genetic correlation between hundreds of traits (Zheng et al., 2017), including GWAS summary statistics for six immune-mediated diseases (Figure 4.1), finding significant correlations between a number of diseases. Whilst such an approach is useful to understand how the overall genetic architecture of two traits might be related, it can be misleading and Bulik-Sullivan et al. (2015) illustrate this using the example of the near zero correlation between both CD with RA even though these diseases are known to share multiple risk loci. This seemingly paradoxical finding occurs because although risk variants are shared between CD and RA (Parkes et al., 2013), these are balanced by variants with opposing effects elsewhere resulting in a negligible net correlation.

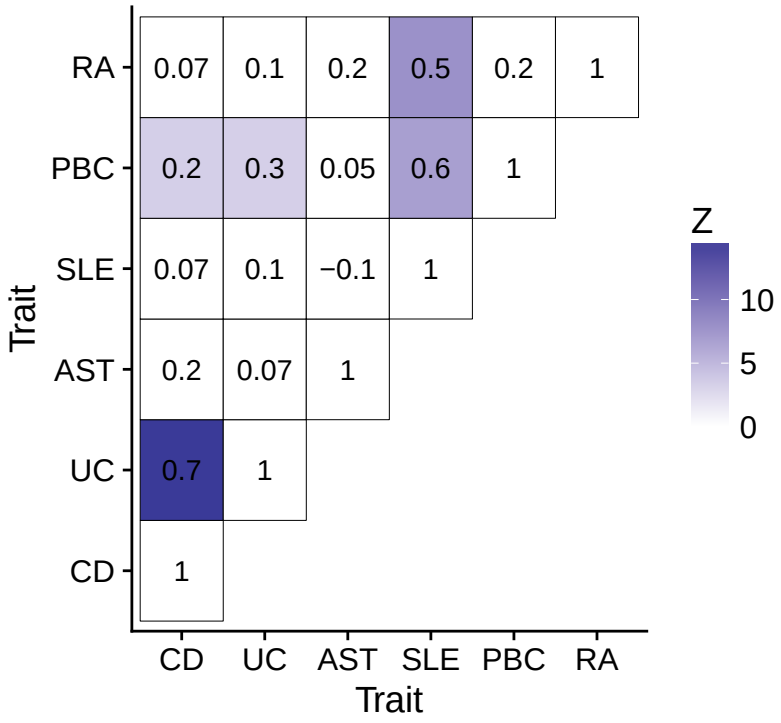


Fig. 4.1 Genetic correlation computed using LD score regression (Bulik-Sullivan et al., 2015). Results were downloaded from LDHub (Zheng et al., 2017) on 06/04/2019 and do not include the HLA region. Pairwise comparisons are labelled with genetic correlation values, and are coloured by Z score if the Bonferroni adjusted p-value < 0.05

A drawback from using genetic correlation to infer shared genetic architectures is that it operates over the whole genome, thus precludes the cataloguing of which risk variants are shared, opposing and distinct across a number of traits presenting substantial challenges for biological interpretation.

Genome-wide multi trait approaches

A more holistic approach, *disPCA* developed by Chang and Keinan (2014), uses principal component analysis (PCA) in order to summarise a matrix of association scores across 31 GWAS datasets. Firstly, for each gene, a summary association statistic, based on genomic distance is computed. The resultant matrix, where rows and columns reflect GWAS studies and genes respectively, is then column mean-centred. This transformed matrix is then suitable for PCA, resulting in loadings for each disease across 31 principal components ordered by the amount of overall variance each explains. Chang and Keinan (2014), chose to concentrate on the first two components, explaining a total of 8.69% of the variance. This analysis demonstrated not only clustering of UC and CD but also SLE with CEL, which was followed up with pathway analysis using the underlying gene loadings, although the results as presented are not convincing.

Overall, this approach has limitations, firstly as I have demonstrated in previous chapters, the assignment of variants to genes on the basis of physical/genetic location is non-trivial and as such it is challenging to know whether the gene-level associations they use as input accurately reflect underlying gene biology. Secondly such gene-level summaries will mask pleiotropy, instead relying on a more diffuse definition of cross-phenotype association that conflicts with convincing results that many loci with shared associations are actually due to distinct causal variants between traits (Fortune et al., 2015). Finally, Chang and Keinan (2014) mention that their approach fails to take into account the direction of effect of underlying causal variants, as such antagonistic and shared effects are treated equally. As discussed in the previous section heterogeneity in effect size and direction between diseases is likely to be pervasive and thus not accounting for this is likely to undermine any results from using *disPCA*.

The approach exemplified by *disPCA* whilst limited, is of interest as it bridges an important methodological gap, using PCA as a dimension reduction technique, to generate interpretable summaries over a genome-wide scale dataset. This invites the exploration of whether such summaries are of utility as lower dimensional composite predictors themselves or whether their underlying loadings, that capture

relationships between SNPs, have further application. The former application, using summaries as covariates in regression based methods, is well described in the literature; for example most standard GWAS analysis pipelines make use of PC scores from ancestral PCA to control for confounding due to population stratification (Price et al., 2006). However, PCA based dimension reduction approaches have also been used to create composite phenotypes, that is lower dimensional summaries, that maximise either the variance or heritability across multiple phenotypes, which are known, *a priori*, to be correlated (Avery et al., 2011; Klei Lambertus et al., 2007). These approaches can provide multiple benefits, not only in terms of maximising the power for association analysis, as the top components by definition maximise the variance between traits, but also in relieving the multiple testing burden when examining multiple related traits (Aschard et al., 2014).

4.3 Basis disease data preparation

In order to create a detailed map of the genetic architecture of immune-mediated disease I created a collection of GWAS summary statistics (Table 4.1) across 10 diseases, from publicly available resources including the GWAS catalogue and ImmunoBase. For inclusion a study needed to fulfil a number of criteria:

Cohort size greater than 10,000 samples: The uncertainty attached to an odds ratio estimate depends on the total number of cases and controls analysed. Whilst arbitrary, I felt that a minimum cohort size of this magnitude allowed the inclusion of cohorts across a wide spectrum of IMD whilst filtering out studies that might lack power and be unlikely to contribute to the proposed approach.

Summary GWAS data availability: The approach requires that estimates of the odds ratios and their significance be available genome-wide. This precludes the inclusion of studies that, for example, present only summary statistics for the genome-wide significant loci.

Study performed on cohorts with European ancestry only: Although a population agnostic view of the genetic architecture of complex disease is favoured (DIAGRAM Consortium et al., 2014; Liu et al., 2015; Okada et al., 2014) there are technical reasons for limiting studies to those of European ancestry in downstream analysis. The main reason is that such a filter

maximises the number of traits that can be studied but also minimises the effects of population stratification on downstream analysis. This is because the contribution to the difference in the estimate of the log odds ratio at a given SNP and disease cohort is less likely to be confounded due to gross heterogeneity in allele frequency that are more likely arise to arise when studies with differences in ancestry are considered. Furthermore, limiting to a specific ancestry allows the use of reference genotypes in order to estimate minor allele frequencies and the fine scaled LD architecture between SNPs for which only GWAS summary statistics are available.

Disease	Abbreviation	Cases	Controls
Primary Biliary Cholangitis	PBC	2,764	10,475
Systemic Lupus Erythematosus	SLE	4,036	6,959
Coeliac Disease	CEL	4,533	10,750
Primary Sclerosing Cholangitis	PSC	4,796	19,955
Type 1 Diabetes	T1D	5,913	8,829
Multiple Sclerosis	MS	9,772	17,376
Crohns' Disease	CD	12,194	28,072
Ulcerative Colitis	UC	12,366	33,609
Rheumatoid Arthritis	RA	14,361	43,923
Asthma	asthma	19,954	107,715

Table 4.1 IMD studies used to construct basis. References for the studies are as follows; PBC (Cordell et al., 2015), SLE (Bentham et al., 2015), CEL (Dubois et al., 2010), PSC (Ji et al., 2017), T1D (Cooper et al., 2017), MS (IMSGC et al., 2011), CD (de Lange et al., 2017), UC (de Lange et al., 2017), RA (Okada et al., 2014) and asthma (Demenais et al., 2018).

In future sections I refer to the logarithm of the sample estimate of the odds ratio, $\hat{\beta}$, as this has the convenient property that $\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}})$. For some studies $\sigma^2_{\hat{\beta}}$ was missing, but can be calculated using

$$\sigma^2_{\hat{\beta}} = \left(\frac{\hat{\beta}}{\Phi^{-1}(p/2)} \right)^2, \tag{4.3}$$

where p is the p-value of the association and Φ^{-1} is the normal quantile function. Throughout the text I make liberal use of the abreviations defined in Table 4.1 using the term 'basis diseases' to refer to them collectively. Finally, I use the term 'basis' as a reference to the resultant eigenvector space from the application of PCA to 'basis diseases' and a synthetic control trait (Section 4.4).

4.3.1 UK10K as a reference genotype dataset

The employment of a reference genotype dataset as a foundation for downstream analyses has multiple benefits and I selected the set of genotypes made publicly available as part of the UK10K project (project consortium, 2015). This study has made available genotypes, identified through whole genome sequencing, for 3,781 healthy individuals from two British cohorts of European ancestry, namely the Avon Longitudinal Study of Parents and Children (ALSPAC) and TwinsUK.

The reasons for this choice were because of its size, coverage, availability and ancestral relevance to the immune disease studies to be analysed. One key benefit is that the challenge in harmonising effect alleles across input studies is ameliorated if this is carried out using a set of reference genotypes for which an accurate estimate of control allele frequencies is available.

4.3.2 SNP selection

The main criteria for selecting SNPs was data availability, the proposed framework requires that GWAS summary data are available for each SNP across all basis diseases. The studies selected encompass a relatively large time interval, in the context of GWAS, and this is reflected in the SNP coverage across studies (Figure 4.2a). This resulted in a set of 294,573 SNPs for which the relevant data was available across all 10 traits. I found the category of SNPs found in only one study was most frequent, driven predominantly by rare SNPs from the Bentham et al. (2015) study of SLE (Figure 4.2b). A second peak of sharing at six studies, reflected the fact that SLE, UC, CD, T1D, RA and PSC studies were all imputed using 1000 genomes data. This exposes the trade off in covering more diseases at the expense of reduced SNP coverage. I decided that maximising the number of basis diseases was of greater value than a denser SNP map based on the fact that the underlying genotyping platforms use a tag SNP approach (Johnson et al., 2001) in order to capture the most genetic diversity by using as few assays/SNPs as possible. A sparser SNP map that still allows signal detection, albeit at a lower resolution is an acceptable trade off if this increases disease coverage, leading to a richer basis.

I excluded SNPs within the HLA region(chr6:20-40Mb) due to its long and complex LD structure. Furthermore it is known to harbour SNPs that have a profound involvement in IMD susceptibility; such large effects have the capacity

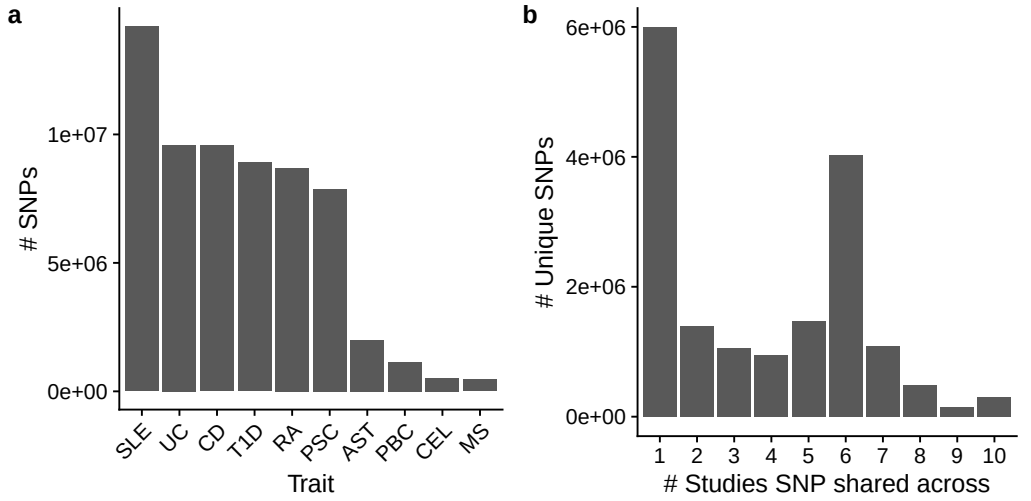


Fig. 4.2 SNP intersection across basis diseases. **a)** Total number of SNPs across all studies for which summary statistics were available. **b)** Sharing of SNPs across studies.

to overwhelm signal due to loci with more modest effect, potentially limiting my ability to interpret downstream analyses.

Finally, I filtered the resultant list of SNPs to remove uncommon or palindromic [1] SNPs (MAF<1%) using the UK10K reference genotyping resource described in section 4.3.1 to obtain a final list of 280,651 SNPs, I refer to this set of SNPs, used for all downstream analysis as 'basis SNPs'. Given the filtering steps employed non-autosomal and mitochondrial SNPs were implicitly excluded as these results were missing from a majority of input GWAS studies and so were not taken forward in the initial SNP study overlap step (Figure 4.3).

Having filtered all 10 studies to include only basis SNPs, I next undertook harmonisation of effect alleles across all studies. This involves adjusting odds ratios such that effect alleles for all studies are aligned with respect to the same allele. The choice of what allele to use for alignment is arbitrary and rather than picking a random input study for such a task I instead aligned all studies to the common UK10K reference set. My main motivation for doing this was to facilitate the integration, at a later stage, of UKBB data, which uses the UK10K reference genotypes for imputations and is therefore automatically harmonised if such a scheme is employed.

[1] These are SNPs such as A/T, for which forward and reverse stranded orientations appear identical. This can result in the miss-assignment of risk alleles undermining any efforts to harmonise effects across multiple studies.

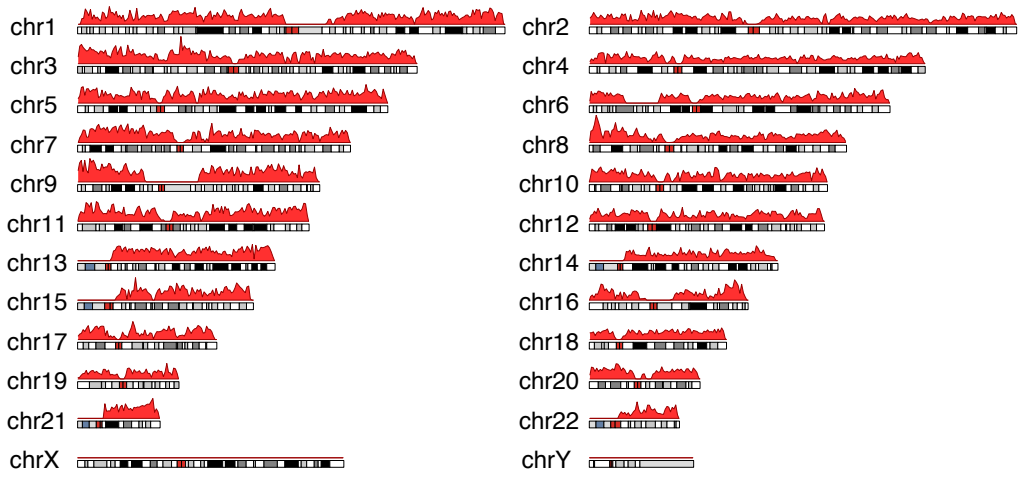


Fig. 4.3 Basis SNP genome coverage.

4.4 Creating a PCA 'basis'

In the previous section I described the preparation of a set of harmonised GWAS summary statistics across ten immune-mediated diseases. I next focused on developing a framework to provide a lower dimensional summary of the similarities and differences in the genetic architectures across traits using principal component analysis (PCA). The output of PCA is a 'basis', a transformation of the input data into a new coordinate system, where each principal axis or component is orthogonal.

To do this I used the aligned datasets to construct a $10 \times 280,651$ matrix of $\hat{\beta}$, where rows and columns reflect diseases and SNPs respectively. I added to this matrix an additional row of zeroes, that constituted a synthetic 'control' disease, modelling a situation where there is no association across any SNPs. I included this 'control' disease in order to provide a reference point or 'baseline' for downstream comparative analysis. I use the term $\mathbf{M}$ to refer to this $11 \times 280,651$ matrix in subsequent sections.

4.4.1 PCA basis creation using $\hat{\beta}$

As described in Chapter 1, PCA, is a relatively simple method, that can be used to construct importance ranked summaries or components of high-dimensional matrices. For moderately sized matrices, such as $\mathbf{M}$, I was able to use the inbuilt R function `prcomp`, in order to conduct PCA. One important consideration prior to PCA, is whether to adjust variables to have unit variance before analysis, to prevent variables of different scales dominating the resultant PCA basis. For a

disease where the underlying study has a large sample size, $\hat{\beta}$ more closely follows the true value of β for that trait, as $\sigma^2_{\hat{\beta}} \propto \sqrt{\frac{1}{N}}$. By scaling diseases to have unit variance we effectively remove this advantage for larger, better powered GWAS studies which is not desirable. I therefore chose not scale **M** prior to PCA. Another consideration is whether to mean centre $\hat{\beta}$ across columns, I chose to do this as one of the reasons for using $\hat{\beta}$ as input is that under the null hypothesis of no association, $\mathbb{E}(\hat{\beta}) = 0$, and such a transformation enforces this.

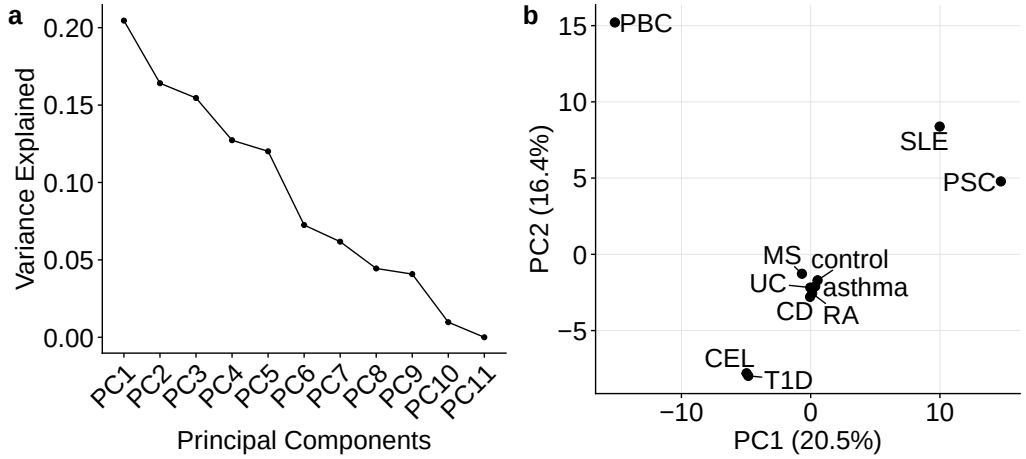


Fig. 4.4 PCA basis using $\hat{\beta}$. **a**) A scree plot of the variance explained by each component. **b**) A biplot of the first two PCs explaining approximately 37% of the variance in $\hat{\beta}$ across all traits examined. Points represent PC scores for a trait (Trait label abbreviations are defined in Table 4.1).

The resultant PCA of **M** resulted in 11 principal components (PCs) where 50% of the variance in $\hat{\beta}$ between diseases was explained in the first three PCs (Figure 4.4a). For this naive $\hat{\beta}$ basis (Figure 4.4b), I found that PC1 described a PBC-SLE/PSC axis, whereas PC2 separated PBC from CEL/T1D, with all other traits clustering around control. Whilst a biplot is a useful summary, it only captures the total variance explained by the first two PCs (37% variance explained). An alternative analysis is to convert the matrix of PC scores across all eleven PCs and traits, into a Euclidean distance matrix. Let p_i and q_i be the PC scores for the i^{th} PC for two traits, the Euclidean distance between them across n PCs is

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \tag{4.4}$$

This distance matrix can then be analysed using an agglomerative hierarchical clustering methods to look for groups of diseases, encompassing all PCs, that are

closest in 'basis' space and therefore share a similar genetic architecture. I chose to use the 'complete linkage' method, where clusters are defined by the farthest neighbour with distances derived from Equation 4.4, as implemented in the R command `hclust`.

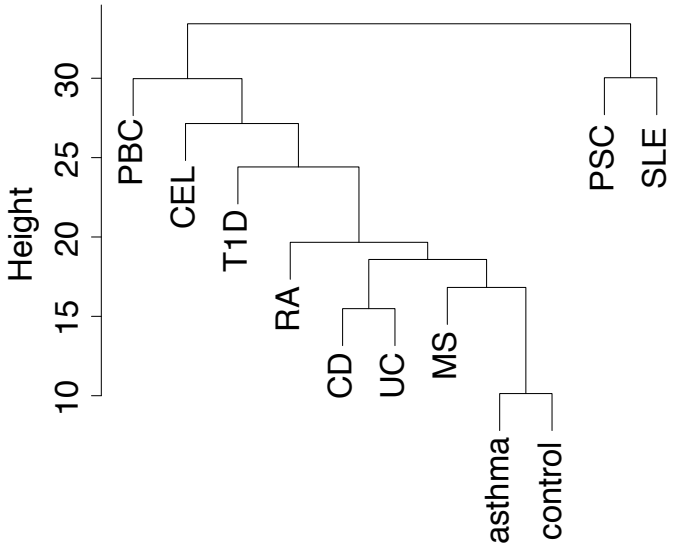


Fig. 4.5 Complete linkage hierarchical clustering of $\hat{\beta}$ PC scores across 11 traits.

This resulted in two main clusters that separated PSC and SLE from the rest of the traits (Figure 4.5), mostly because of the large effect of PC1. The reported genetic correlation between PSC and UC (Ji et al., 2017) was not evident in this clustering, a result that is in conflict with clinical observations that around 75% of individuals with PSC go on to develop IBD. This coupled with the poor discriminatory performance between control and disease for the top two PC components, explaining over a third of the variance, undermined the utility of a basis constructed from raw $\hat{\beta}$.

4.4.2 PCA basis creating using $\hat{\gamma}$

In the previous section I used PCA on $\mathbf{M}$, a matrix of aligned $\hat{\beta}$, of diseases in order to define a subspace, such that diseases with similar $\hat{\beta}$ are proximal when projected into this basis. One source of uninformative variance between the columns of $\mathbf{M}$ is related to the allele frequency, f of a SNP, which relates to the standard error of $\hat{\beta}$ thus

$$\sigma_{\hat{\beta}} = \sqrt{\frac{1}{2N}}\sqrt{\frac{1}{\hat{f}} + \frac{1}{1-\hat{f}}}, \tag{4.5}$$

under the null hypothesis that $\beta = 0$, where $\hat{f}$ is an estimate of the minor allele frequency in the cohort for a given basis SNP. Equation 4.5 is derived from the 2×2 table that underlies an odds ratio, where the standard error using the labelling defined in Figure 1.1 is $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. Under the assumption that $\hat{f}$ is the same across basis traits, which is not unreasonable given that they are all of European ancestry, this source of variance between studies should be removed as it could obscure variance due to genuine differences in genetic architecture between basis diseases. By rearranging equation 4.5, it is possible to estimate the component of σ_β that is due to allele frequency from the available summary statistics by

$$\hat{\sigma}_f = \sqrt{2N}\sigma_{\hat{\beta}} = \frac{1}{\sqrt{\hat{f}(1-\hat{f})}}. \tag{4.6}$$

Such an approach generates $\hat{\sigma}_f$ for each SNP and basis disease pairing, and I chose to take forward the mean $\hat{\sigma}_f$ per SNP using this to adjust $\hat{\beta}$ for each basis disease by employing

$$\hat{\gamma} = \frac{\hat{\beta}}{\hat{\sigma}_f}, \tag{4.7}$$

resulting in a new matrix of $\hat{\gamma}$ which I used to compute a new basis (Figure 4.6).

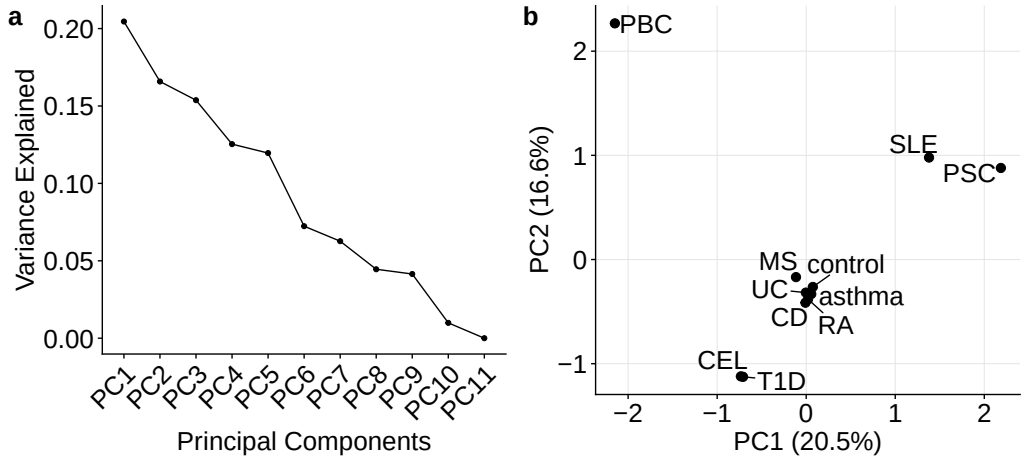


Fig. 4.6 PCA basis using $\hat{\gamma}$. **a**) A scree plot of the variance explained by each component. **b**) A biplot of the first two PCs.

I found that the basis generated, whilst on a different scale to that generated using $\hat{\beta}$ was very similar resulting only minor differences in the variance explained across PCs. This is to be expected as the values of $\hat{\gamma}$ whilst different between SNPs are constant across input traits. This results in a rescaled basis space where PC scores for traits are simply rescaled but their relative positions remain the

Self Reported Disease	Label	No. Cases	No. Controls
Asthma	bb_asthma	39,049	298,110
Rheumatoid arthritis	bb_RA	3,730	333,429
Ulcerative colitis	bb_UC	1,795	335,364
Malabsorption Coeliac disease	bb_CEL	1,452	335,707
Multiple sclerosis	bb_MS	1,228	335,931
Crohns Disease	bb_CD	1,032	336,127
Systemic lupus erythematosis	bb_SLE	366	336,793
Type 1 diabetes	bb_T1D	286	336,873

Table 4.2 UK BioBank self-reported phenotypes[https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=20002] matching diseases from which the basis was constructed.

same. For GWAS where MAF are greater than 0.1, the standard error due to sample size is orders of magnitude smaller than $\hat{\gamma}$. This means that, in practice, the influence of MAF on $\sigma_{\hat{\beta}}$ is small for the cohort sizes and allele frequencies usually encountered in GWAS. However in this context which is motivated by a wish to consider GWAS performed on small cohorts its derivation and inclusion in the scaling is justified.

4.5 Evaluating basis performance

One approach to examining the relevance of the basis space developed in the previous section is to interrogate its ability to predict the PC scores for basis traits from an alternative source of GWAS summary statistics.

One source of summary statistics over thousands of traits is the UK Biobank (UKBB) (Sudlow et al., 2015). This large cohort has collected both genetic and phenotypic data over a large range of traits for approximately 500,000 individuals. Performing high quality GWAS across such a large cohort for so many traits is technically challenging. To overcome this the Neale laboratory have created software tools such as HAIL (https://github.com/hail-is/hail) and PHESANT (Millard et al., 2017), which they have used to generate and release publicly the summary statistics for over 11,000 UKBB traits (http://www.nealelab.is/uk-biobank/). From this resource I selected eight self-reported diseases (SRD) for which there was a matching basis trait (Table 4.2).

In general these SRD GWAS contained an order of magnitude fewer cases, resulting in much less powered GWAS, for example, self-reported T1D covers only 286 cases, compared to the 5,913 included in the basis. The exception is asthma

which contained nearly double the number of cases included from Demenais et al. (2018). It is important to note that these diseases, are by definition, self-reported phenotypes, and as such are likely to be significantly 'softer' than those from which the basis is constructed. By this I mean case/control status is not necessarily driven by a clinical diagnosis which can result in a misclassification of both case and control status which further undermines power. I hypothesised that if the basis reflected true differences in the genetic architecture between input traits, that these matched SRD, when projected into basis space, should be closer to their counterparts than other diseases.

4.5.1 Linear regression coefficient conversion to the odds ratio scale for a binary trait

For computational efficiency, the Neale laboratory have employed a linear regression approach to all data considered, even if the trait under consideration is binary. Thus, in order for me to project these data onto the basis, summary statistics needed to be transformed to the matching log odds ratio scale.

The UKBB summary statistics downloaded from the Neale Laboratory compendium include the total number of samples considered N, as well as N_1 the total number of affected samples (i.e. exhibit the particular self-reported binary phenotype). Also reported are $\sum_i^N X_i Y_i$ (i.e. $X^T Y$) and $\sum_i^N X_i$ where i indexes individuals included in the study of each trait. These quantities for each SNP, are sufficient to compute an odds ratio, under the assumption of Hardy-Weinberg equilibrium under an additive model. In the case of a binary trait where $Y \in \{0,1\}$, with 0 and 1 representing whether a given sample is a control or case respectively, the allele frequency in cases is

$$\hat{f}_{\text{case}} = \frac{\sum_i^N X_i Y_i}{2N_1}. \tag{4.8}$$

Conversely, the allele frequency in controls is

$$\hat{f}_{\text{ctrl}} = \frac{\sum_i^N X_i - \sum_i^N X_i Y_i}{2(N - N_1)}, \tag{4.9}$$

and the estimated odds ratio is therefore

$$\frac{\hat{f}_{\text{case}}(1 - \hat{f}_{\text{ctrl}})}{\hat{f}_{\text{ctrl}}(1 - \hat{f}_{\text{case}})}. \tag{4.10}$$

4.5.2 Comparison of $\hat{\gamma}$ basis PC scores with matched UKBB self-reported projections

Using this method I computed $\hat{\beta}$s across all basis SNPs for each of the self-reported traits using the basis values of $\hat{\sigma}_f$ and equation 4.7 to convert these to $\hat{\gamma}$. Given that the basis was constructed with $\hat{\gamma}$ aligned with the UK10K reference, and UKBB genotypes have been imputed using the same reference alleles were already aligned. Finally I used R function `predict`, in order to project each of the selected UKBB SRDs (Table 4.1) onto the $\hat{\gamma}$ previously described (Section 4.4.2). The outcome of this is a 9×11 matrix of PC scores, where rows reflect projected traits and columns the total number of PCs described by the basis.

In order to investigate whether there was support for projected SRDs to have, overall, similar PC scores to their basis counterparts I used complete linkage hierarchical clustering of the matrix of PC scores of both basis and projected diseases (Section 4.4.1).

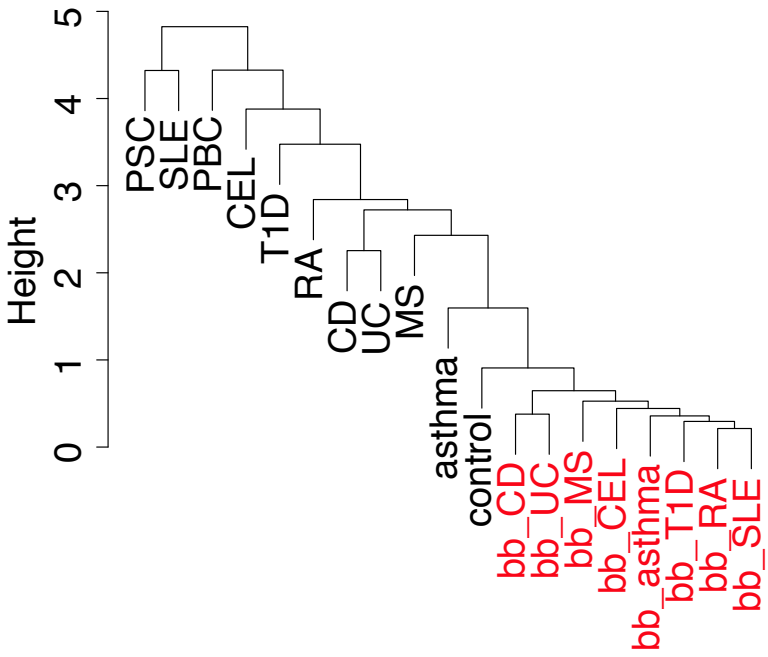


Fig. 4.7 Complete linkage hierarchical clustering of $\hat{\gamma}$ PC scores across 11 traits (black) and matched UKBB SRDs (red).

I found that such an approach did not support my hypothesis that when projected, matched basis and UKBB SRD would occupy similar positions in basis space (Figure 4.7). Instead, all projected UKBB SRD clustered together rather than with their corresponding matched basis trait. Furthermore the closest clustered

basis trait was the synthetic 'control' trait. These observations, taken together with the fact that basis trait clustering does not reflect previous studies on shared genetic architectures of IMD (Cho and Feldman, 2015; Ji et al., 2017; Márquez et al., 2018), suggests that the basis as constructed is still overwhelmed by sources of variance between traits not related to their genetic architecture.

4.6 Development of a Bayesian shrinkage method

In the previous sections I described the creation of a basis in order to provide a low dimensional representation of the shared and distinct genetic architectures of ten immune-mediated diseases. However, evaluation of the resultant basis showed that it was not reflective of known disease overlap. The main reason for this is that the current proposal uses as input, log transformed odds ratio estimates, with no regard to their significance. Whilst there are arguments, in the form of the omnigenic model (Boyle et al., 2017), that a large number of variants are likely to both directly and indirectly influence a trait, most will have such small effect sizes, such that that large sample sizes would be required for their accurate estimation. This means that a vast majority of odds ratios estimates for SNPs included in the basis are 'noisy' and have large standard errors associated with them. It seems reasonable that such a large source of variance arising from both technical and stochastic processes would be expected to overwhelm the basis, thus obscuring true associations driving similarities and differences between the genetic architecture of diseases.

One way of taking account of this stochastic variance is to apply a threshold and only select SNPs exhibiting significant association in one or more traits. Such an approach whilst attractive due to its apparent simplicity has several drawbacks. Leaving aside the Winner's Curse effect (Section 2.2.1), the foremost challenge is to develop a suitable strategy by which to select such a threshold. If for example we select only genome-wide significant variants ($p < 5 \times 10^{-8}$), those variants that are below but close to this threshold threshold will be unable to contribute to basis generation even though they may contain useful information. Another difficulty is how to adequately deal with LD differences across the genome. If a locus is in strong LD, then by thresholding it is likely that more SNPs are included in the basis compared to a locus in weaker LD. This means that loci with high LD will have greater influence on the basis generated, thus creating a bias. To summarise,

any proposed method should not rely on thresholding and also intrinsically take into account the differential amounts of LD across the genome.

4.6.1 Method description

One approach that fulfils these criteria is the Bayesian single causal variant fine-mapping approach introduced in previous chapters that summarises, in the context of an arbitrary genomic region, the posterior probability for a variant to be causal. A key facet of this approach is that posterior probabilities are jointly modelled across all variants within a the genomic region considered. This property is useful as if regions are selected in such a way that they are approximately LD independent from each other, the effect of LD can be largely mitigated. Such a method, when combined over multiple diseases, can be used to assign individual SNP weights, that can then be used to rescale input odds ratios. These rescaled odds ratios, weighted or shrunk in proportion to their disease relevance, can be used to generate a basis that holds promise in overcoming some of the challenges discussed in the previous section.

To describe the implementation of such a weighting scheme in detail, consider a genomic region, r that is approximately LD independent from neighbouring regions. As discussed in the previous chapter if for a particular trait, we have association p-values, $\hat{f}$ and $\hat{\beta}$ for a set of SNP with r, then we can compute a single causal variant posterior probability (sCVPP) for each SNP (Section 1.3.5). The next challenge is integrating sets of sCVPP across multiple diseases. Let i and j index a matrix of sCVPP across s SNPs and d, diseases, respectively, within region r. I first compute a disease specific weighting such that

$$\alpha_j = \sum_{i=1}^{s} \text{sCVPP}_{ij}. \tag{4.11}$$

To combine across diseases I compute a disease weighted average for the i^{th} SNP sCVPP within r as

$$h_i = \frac{\sum_{j=1}^{d} \alpha_j \text{sCVPP}_{ij}}{\sum_{j=1}^{d} \alpha_j}, \tag{4.12}$$

which once multiplied by $\hat{\gamma}$ leads to $\hat{\gamma}^{'}$, a vector of shrunk $\hat{\gamma}$ of length s for a target disease.

For ease, given the invariant nature of both σ_f and h across diseases, in future sections I refer to h_i', as the combined weighting of the i^{th} SNP such that

$$h'_i = \frac{h_i}{\sigma_f}. \tag{4.13}$$

I illustrate this process by using the example of seven IMDs (ommiting three diseases with no association across the region for clarity) across the 2q33.2 region (Figure 4.8). In this region there is evidence for a strong association proximal to the gene *CTLA4* in both RA and T1D, however LD within the region makes identifying further associated SNPs in other diseases challenging. By computing sCVPP, I found some evidence for putative causal variants in other diseases that are distinct from RA and T1D. In order to combine these sCVPP into a single weight I consider two things. Firstly, for a given disease the posterior probability that the 2q33.2 region being considered contains a causal variant[2], which is reflected in α_j. Secondly, the disease weighted (using α) contribution of a particular SNP to the total posterior probability for the region, across all diseases. In the example this results in most of weight being applied to the SNP proximal to *CTLA4* previously identified in T1D and RA, with most other variants receiving almost no weight. Importantly, putative causal variants in the other diseases such as PSC and CEL are non-zero weighted, but because there is less evidence for these both at the sCVPP level and across diseases their weight is attenuated. The application of these weights results in both a relative down weighting of SNPs unlikely to be involved in disease susceptibility and a concomitant up weighting on SNPs where evidence exists for their involvement in disease.

The scheme as described applies only to an approximately LD independent region r, to apply it genome-wide I use a similar approach to that described in the previous chapters. Briefly, I split the genome into approximately independent LD blocks using recombination data from the international HapMap project (International HapMap Consortium et al., 2007). I then compute weightings separately for each block, concatenating these in order to obtain the complete set of weightings across all SNPs to be included in the basis.

Whilst the underlying approach was developed for genetic fine-mapping it is important to note that this is a separate goal from the weighting metric that we seek. The density of SNPs considered precludes accurate fine-mapping and as illustrated in the previous chapter we rely on their being up to one causal variant in a given LD block. In contrast, here we seek a weight that best captures how likely a given SNP within the region is relevant across all basis diseases being considered. By performing PCA we summarise information across all basis variants, and thus

[2]Under the assumption that the causal variant is included in the set of SNPs being considered

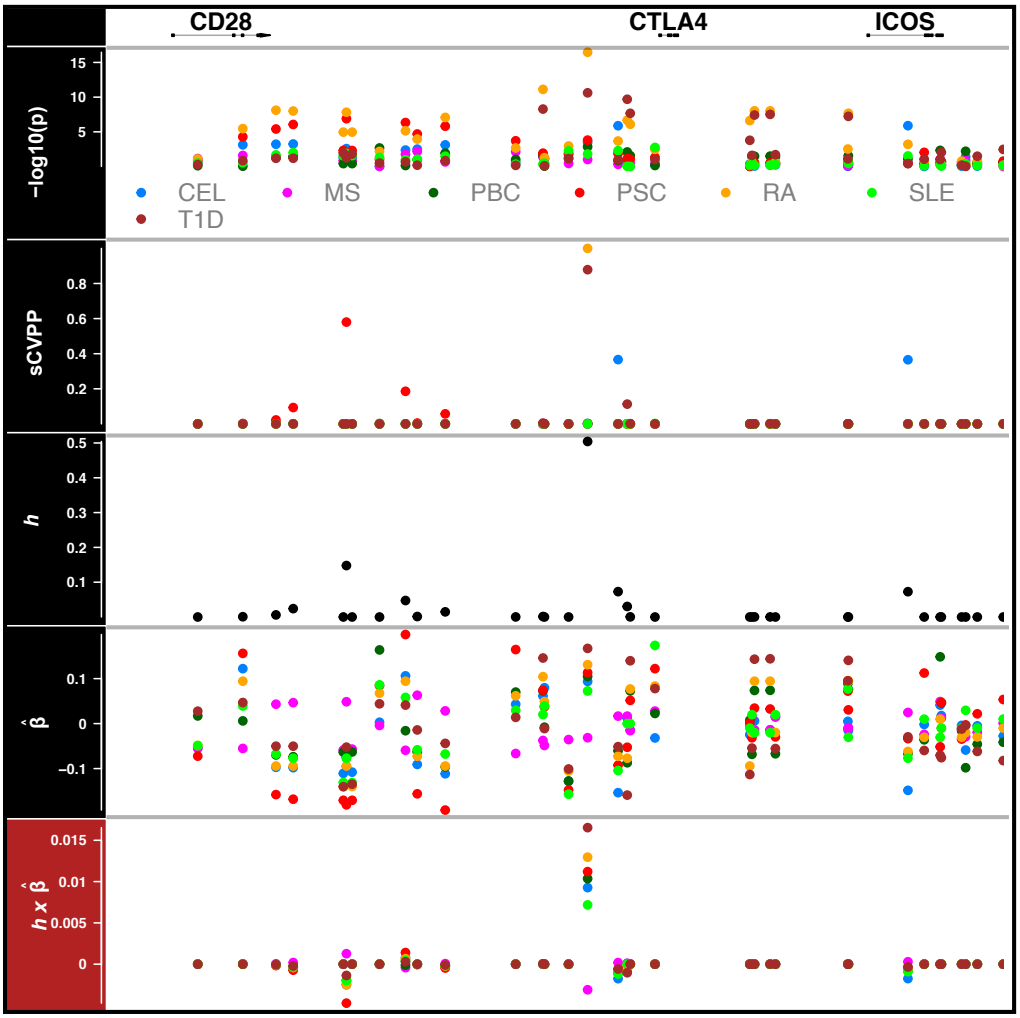


Fig. 4.8 An example of the $\hat{\beta}$ weighting scheme across seven selected IMDs in the 2q33.2 region. The top stanza shows gene positions for GRCh37 as obtained from Ensembl BioMart. $-\log_{10}(p)$ - are the transformed p-values for each study coloured by disease as follows: blue - Coeliac disease (CEL), pink - Multiple sclerosis (MS), darkgreen - Primary biliary cholangitis (PBC), red - Primary sclerosing cholangitis (PSC), orange - Rheumatoid arthritis (RA), lightgreen - Systemic lupus erythematosis (SLE) and brown - Type 1 diabetes (T1D). This colouring scheme is repeated for all subsequent relevant stanzas. sCVPP - single casual variant posterior probabilites for each disease. h - final cross disease weights. $\hat{\beta}$ - The raw estimates of the log(Odds Ratio) across disease. $h \times \hat{\beta}$ - weight transformed log(Odds Ratio).

as long as violations of underlying fine mapping assumptions are not systematic the weighting is justified as these will be distributed over the thousands of regions and the relatively large number of SNPs being considered.

4.6.2 Shrinkage evaluation

Using the basis diseases I generated weightings across all 280,651 input SNPs. The range of $\mathbf{h}$ was between 1.2×10^{-8} and 0.21 with 99% of weights being below 0.05. I applied these weightings to $\mathbf{M}$, the matrix of $\hat{\beta}$, by multiplying each element by its corresponding h' and used the resultant matrix to compute a basis as previously described in section 4.4.1.

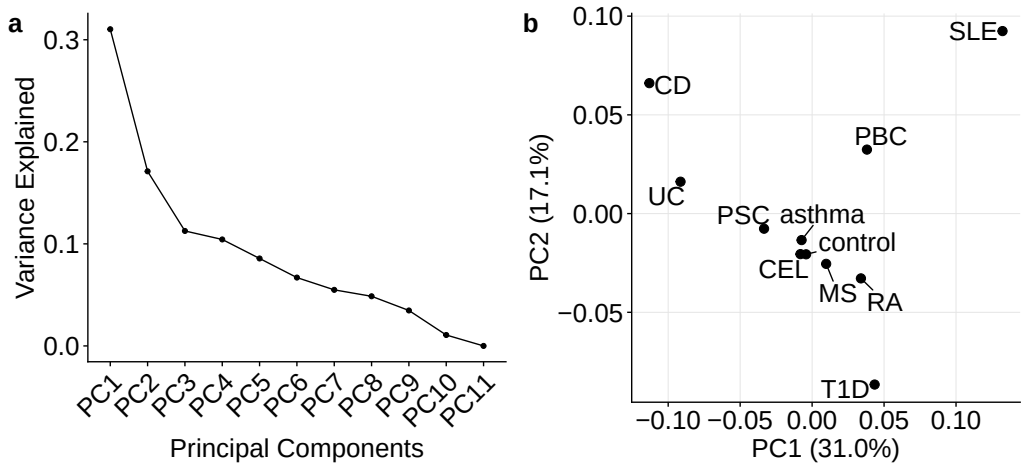


Fig. 4.9 Basis PCA using shrunk $\hat{\gamma}$. **a**) A scree plot of the variance explained by each component. **b**) A biplot of the first two PCs.

For the resultant basis I found that a much larger proportion ($\approx 60\%$) of the variance explained was found in the first three principal components compared with previous basis that used $\hat{\gamma}$ as an input (Figure 4.9a). Furthermore these PC scores for the first two input diseases seem to reflect known disease biology. For example PC1 seems to discriminate auto-inflammatory diseases such as CD, UC and PSC from those categorised as more classical autoimmune diseases such as SLE, PBC and T1D (McGonagle and McDermott, 2006).

I evaluated the resultant basis, as before (Section 4.5), using matched UKBB SRD with the proposed Bayesian shrinkage scheme applied prior to projection onto the basis (Figure 4.10).

I found that this basis produced PC scores for projected SRD UKBB traits that were more similar to their basis counterparts than between other diseases. The only exception to this was for RA which appeared to match more closely with the synthetic control trait. One possible explanation for this is that the projected studies are all self-reported and symptoms of RA are more often confused with other conditions such as osteoarthritis than other projected traits.

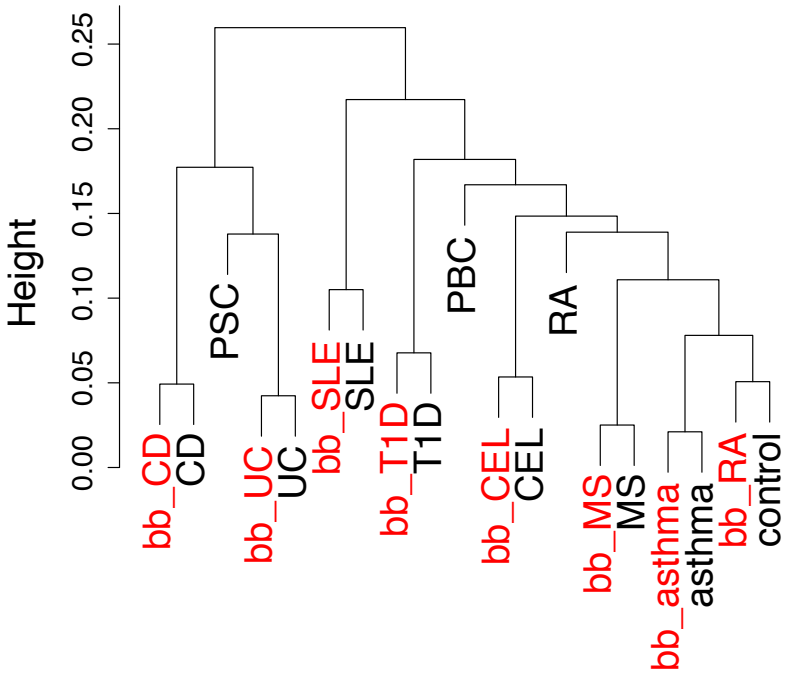


Fig. 4.10 Complete linkage hierarchical clustering of weighted $\hat{\beta}$ PC scores across 11 basis traits (black) and matched UKBB self-reported traits (red).

4.7 Estimating the variance of projected PC scores

In the previous section I proposed a framework, based on a Bayesian weighting method, for generating a basis which was supported by widespread co-clustering of matched basis and UKBB SRD diseases. Whilst PC scores for diseases used to construct the basis are fixed, as these same diseases are used to construct the basis, scores for projected diseases will have an associated sample variance. Efficiently, and accurately estimating this variance is important as it is required in order to make inference about whether individual or sets of projected traits are significantly different from one another.

4.7.1 Analytical variance estimation

In order to consider the variance associated with projecting a trait onto a basis PC, consider the simple case of a single approximately LD independent region containing s variants, indexed by i. To project the i^{th} SNP onto the k^{th} basis PC, requires first taking the product of between $\hat{\beta}_i$ and the corresponding weights (h_i') and PC loading (l_{ik}).

$$q_{ik} = \hat{\beta}_i h_i' l_{ik}. \tag{4.14}$$

When performed across all s variants this results in vector $\mathbf{q}_k = \{q_{1k} \dots q_{s-1,k}, q_{sk}\}$. The variance of q_{ik}, $\text{Var}(q_{ik})$, depends on the variance of $\hat{\beta}$, and h_i' and l_{ik} which are treated as constants such that

$$\begin{aligned} \text{Var}(q_{ik}) &= \left(h_i' l_{ik}\right)^2 \text{Var}\left(\hat{\beta}_i\right) \\ &= \left(\frac{h_i}{\sigma_f} l_{ik}\right)^2 \sigma_N^2 \sigma_f^2 \\ &= \left(h_i l_{ik}\right)^2 \sigma_N^2, \end{aligned} \tag{4.15}$$

where σ_N^2 and σ_f^2 are the component of the standard error of $\hat{\beta}$ due to sample size and MAF respectively. The variance-covariance matrix of $\text{Var}(q_k)$ requires taking into account the linkage disequilibrium between SNPs. To do this I use the fact that the correlation matrix of genotypes, $\mathbf{\Sigma}$, which can be estimated from a set of reference genotypes such as the UK10K, is the same as the variance-covariance matrix of standardised $\hat{\beta}$, $\mathbf{Z}$, where each element is a z-score such that $z = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$ (Burren et al., 2014). Letting $\mathbf{V}_k = \left(\sqrt{\text{Var}(q_{ik})}, i = 1, \dots, s\right)$ the variance-covariance matrix, $\mathbf{X}_k^{s \times s}$ associated with projecting the s variants, onto the k^{th} PC in the simplified region can be approximated as:

$$\mathbf{X}_k = (\mathbf{V}_k \circ \boldsymbol{\sigma}_{\hat{f}})^T (\mathbf{V}_k \circ \boldsymbol{\sigma}_{\hat{f}}) \circ \mathbf{\Sigma}, \tag{4.16}$$

where vector $\boldsymbol{\sigma}_{\hat{f}}$ is the component of the standard error due to allele frequency across s SNPs and $\circ$ denotes element-wise multiplication. The variance associated with the projection onto the k^{th} PC over the v^{th} region is then the the sum over all elements of $\mathbf{X}_k$

$$\text{Var}(\text{P}_{kv}) = \mathbb{1} \mathbf{X}_k \mathbb{1}^T, \tag{4.17}$$

where $\mathbb{1}$ is row vector of 1s of length s.

This result can be extended genome-wide by making the assumption of independence between regions, which is reasonable if the regions have been selected to be approximately LD independent. Then the total variance for a particular PC is simply the sum of the variance for that PC across the total number of LD blocks, r, to be considered,

$$\mathrm{Var}(\mathrm{P}_k) = \sum_{v=1}^{r} P_{kv}. \tag{4.18}$$

Whilst such a scheme takes into account the component of variance due to minor allele frequency, $\sigma^2_{\hat{f}}$ it does not take into account the component due to sample size, σ^2_N. This is beneficial as under the reasonable assumption that study design is fixed across all projected SNPs, this quantity is constant allowing the computation of the variance associated with the projection of any case/control study design. I approximate σ^2_N as

$$\sigma^2_N = \frac{N_1 + N_0}{N_0 N_1}, \tag{4.19}$$

where N_1 and N_0 are the case and control size of the projected study respectively. This means that the variance of projecting an arbitrary trait, t onto the the k^{th} PC is

$$\frac{N_1 + N_0}{N_0 N_1} \times \mathrm{Var}(P_k), \tag{4.20}$$

where N_1 and N_0 are the case and control size of t GWAS respectively. This setup provides the benefit that the value $\mathrm{Var}(P_k)$ needs only be estimated once for a given basis, mitigating the computational overhead associated with having to empirically estimate the variance associated with projecting many traits onto the basis.

4.7.2 Empirical variance estimation

In the previous section I presented an analytical result that makes both implicit and explicit assumptions about LD independence, due to its reliance on the approximation of $\mathbf{\Sigma}$ from a reference genotype set and approximately LD independent genomic regions. In order to validate the result I next developed simulations for evaluating the variance of a projected trait under different conditions.

Simulation under H_0

The first condition I examined was the projection of a synthetic GWAS trait under the null hypothesis of no association. Under this condition, $\mathbb{E}(\hat{\beta}) = 0$, and an estimate of $\hat{\beta}$ for a single SNP can be obtained from $\hat{\beta} = N(0, \sigma^2_{\hat{\beta}})$, where

$$\sigma^2_{\hat{\beta}} \approx \frac{1}{2N}\left(\frac{1}{\hat{f}} + \frac{1}{(1-\hat{f})}\right), \tag{4.21}$$

where N is the simulated sample size under consideration, and $\hat{f}$ is the estimated minor allele frequency obtained from UK10K reference genotype resource. In order to simulate sets of $\hat{\beta}$ under H_0 under differing sample sizes robustly we must incorporate LD. Thus the simulation of $\hat{\beta}s$ for a given LD block involves sampling from a multivariate normal,

$$\hat{\beta} \sim \text{MVN}\left(0, \left(\sigma_{\hat{\beta}}\sigma_{\hat{\beta}}^{T}\right) \circ \mathbf{\Sigma}\right). \tag{4.22}$$

For a given sample size this is done for each LD block, and the results concatenated to create a vector of $\hat{\beta}$ under H_0, which can be re-scaled and projected onto the basis.

Under the assumption that PC scores from this simulation will be distributed normally it is possible to estimate approximately, the number of simulations required for a reasonable empirical estimate of the PC score variance. This involves setting a target variance, σ^2, and then simulating PC scores such that $\text{PC}_{\text{score}} \sim N(0, \sigma^2)$, 1,000 times for each value of n, where $n = \{200, 500, 10^3, 2 \times 10^3, 10^4, 10^5\}$ reflecting the number of simulations performed, and setting $\sigma^2 = 0.1$, a reasonable estimate given the observed basis PC scores (Figure 4.9). At each n I computed the coefficient of variation (CV) $\frac{\hat{\sigma}}{\hat{\mu}}$ where $\hat{\sigma}$ and $\hat{\mu}$ are estimates for the standard error and mean computed over the 1,000 simulations for a given n. CVs had a range from 5% where $n = 200$ to 0.2% where $n = 10^5$. I took forward an $n = 500$ as this gave a CV of 3% that balanced the stability of the variance estimate with the increasing computational burden associated with larger values of n. Using the proposed framework I simulated and projected GWAS summary statistics under H_0 across sample sizes ranging from 1,000 to 15,000 individuals.

Simulation under H_1

The previous simulations are generated under the null hypothesis of no association across any SNPs. In application such a strong assumption will be violated and therefore I also wanted to assess how appropriate both analytical and simulated PC score variance were compared to those obtained from actual GWAS data.

To do this I used imputed genotypes for T1D from WTCCC (Wellcome Trust Case Control Consortium, 2007) supplied by Chris Wallace. I carried out QC, filtering out SNPs with poor imputation scores, violating HWE ($Z_{HWE} > 5$), and not overlapping the set of basis SNPs. This left approximately 286,000 variants across 5,271 individuals (1,929 cases and 3,342 controls). I used this set of variants

to create a basis as described previously (Section 4.6.2) with the T1D trait excluded. This is because additional correlation between $\hat{\beta}$s between a basis and projected traits would be introduced which is not due to a shared disease architecture but instead because of sample sharing. This correlation is non negligible when there is a large degree of sample overlap between traits such as occurs for Cooper et al. (2017) and Wellcome Trust Case Control Consortium (2007) GWASs (Discussed further in Section 4.9.5).

In order to estimate the variance associated with the projection PC scores I employed a bootstrap methodology (Efron, 1979). Let N_1 and N_0 be the number of cases and controls for which the variance of PC scores are to be estimated. In the first step I sample N_1 case and N_0 control genotypes to form a 'source' set of genotypes. To perform a bootstrap estimate of the PC variance for a particular case/control configuration I sample *with replacement* N_1 case and N_0 control genotypes from this 'source' set, which I use to conduct a case/control GWAS (using 1958 birth cohort codings for regional locations as a covariate). Finally the resultant $\hat{\beta}$ are projected onto the basis in the usual fashion to derive PC scores which are stored. I repeated both steps 500 times for a given value of N_0 and N_1 allowing a robust estimation of the variance PC scores under different GWAS case and control size scenarios. Unlike the H_0 framework I was limited by the underlying WTCCC study size in the configurations that can be assessed, I therefore assessed matched numbers of cases and controls for sample sizes of 500,1000 and 1900.

A further source of variance not related to disease genetic architecture might be introduced by population substructure. Whilst the bootstrap above does incorporate geographical birth location as a covariate in order to mitigate this, the sampling regime as described has the potential to preferentially select cases or controls from a particular region. To obviate this I performed the bootstrap a second time making sure that geographic location proportions amongst cases and controls were preserved after sampling.

4.7.3 Variance estimate evaluation

The results of simulations and bootstraps indicate that the analytical estimates of the variance of PC scores were well calibrated (Figure 4.11). Empirical estimates of the variances under H_0, both H_1 with and without population stratification were aligned with analytical estimates and well within 95% confidence intervals supporting their utility in downstream analysis.

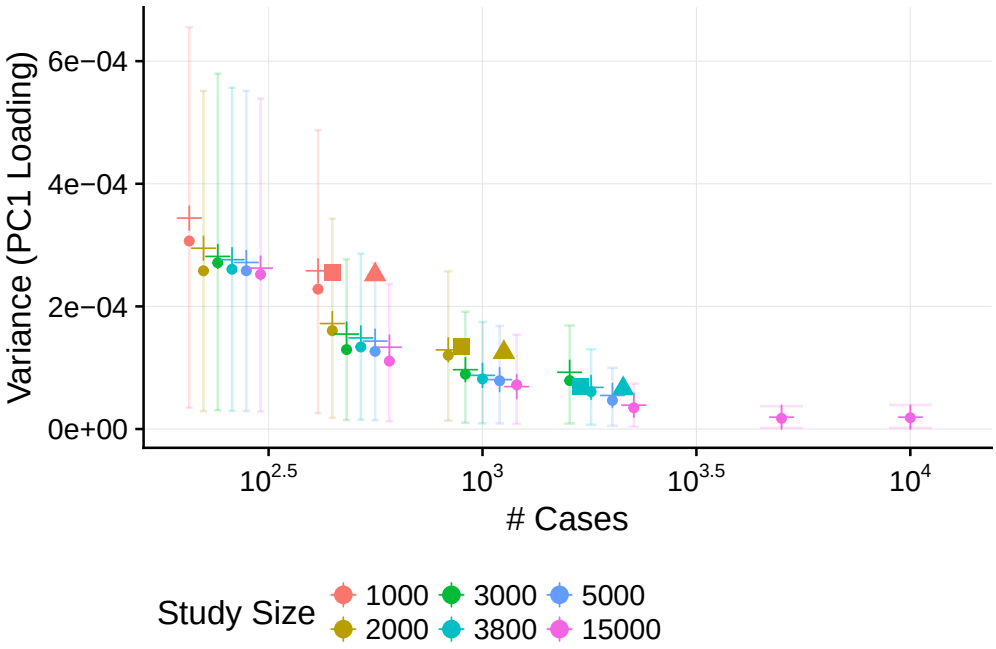


Fig. 4.11 Variance estimation of PC score method comparison. Crosses represent analytical variance estimates under H_0 (Section 4.7). Points represent empirical computation of variance over 500 simulations of projected summary stats under H_0 (Section 4.7.2). Triangular points represent projection variances from 500 bootstraps at different sample size configurations under H_1 (Section 4.7.2), with squares representing where geographical location has been taken into account in the sampling strategy.

4.7.4 Assessing the significance of trait projections

Having a reliable method for estimating the variance associated with trait projection into the basis space is useful as it can be used to assess whether a particular PC score for an axis is significantly different from the control trait, where no associations are expected. In subsequent sections I use $\delta_{kt} = P_{kt} - P_{k,\text{control}}$ to centre projections about control where, as before k indexes the number of PCs. After such a transform, a standard normal Z-score can be computed for the the k^{th} PC as

$$Z_{kt} = \frac{\delta_{kt}}{\sqrt{\text{Var}(P_k)}}. \tag{4.23}$$

4.8 Annotation of basis principal components

I next sought to better understand the biology that might underlie individual basis PCs. One mechanism for doing this is to project GWAS summary statistics for a

diverse range of traits onto the basis and catalogue PC and trait combinations that are significantly different from the 'control' basis trait PC score.

4.8.1 Projection of UKBB self-reported trait GWAS

One source of traits is the previously mentioned the UKBB (Sudlow et al., 2015). In addition to the seven self-reported traits already projected (Section 4.5) there are further traits with binary outcomes for which summary statistics are available from the Neal laboratory compendium. I chose to take forward three main categories covering self-reported medical conditions (UKBB code 20002, $n = 282$) cancer (UKBB code 20001, $n = 28$) and treatment/medications[3] (UKBB code 20003, $n = 542$), which are all self-reported.

I downloaded summary statistics for all of the categories mentioned from the Neale Laboratory compendium, filtering to obtain summary statistics for SNPs included in the basis. After converting to the odds ratio scale (Section 4.5.1) and applying weights (Section 4.6) I projected all 854 traits onto the basis. I used the analytical variance estimations (Section 4.7) of PC score projections to assign significance to combinations of traits and principal components. To select combinations for further study I used the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) to adjust p-values for multiple testing over all trait and PC combinations, selecting those with a false discovery rate (FDR) less than 5% for downstream analysis. In total across all traits I took forward 210 significant trait-PC combinations, compassing 108 unique traits in order to annotate basis components (Figure 4.12). I decided to exclude PC11 from downstream analysis, due to the small amount of overall variance it explains, and the fact that this likely composed of a large proportion of stochastic variance.

PC1 This component is effective at discriminating between autoinflammatory diseases, including CD and UC with negative δ values, and those of a more classical autoimmune phenotype such as RA and SLE whereδ is positive (Figure 4.13). Gastroenteritis/Dysentry has the most positive δ values, however it is likely that this is a false positive result, given that the number of cases is small (n=104), and such a result contradicts clinical observations that link gastroenteritis to an increased risk of developing IBD (García Rodríguez et al., 2006). Self reported medication δ values tend to correlate

[3]This category contains data on any regular treatments taken weekly, monthly, etc. It does not include short-term medications (such as a 1 week course of antibiotics) or prescribed medication that is not taken, or over-the-counter medications, vitamins and supplements.

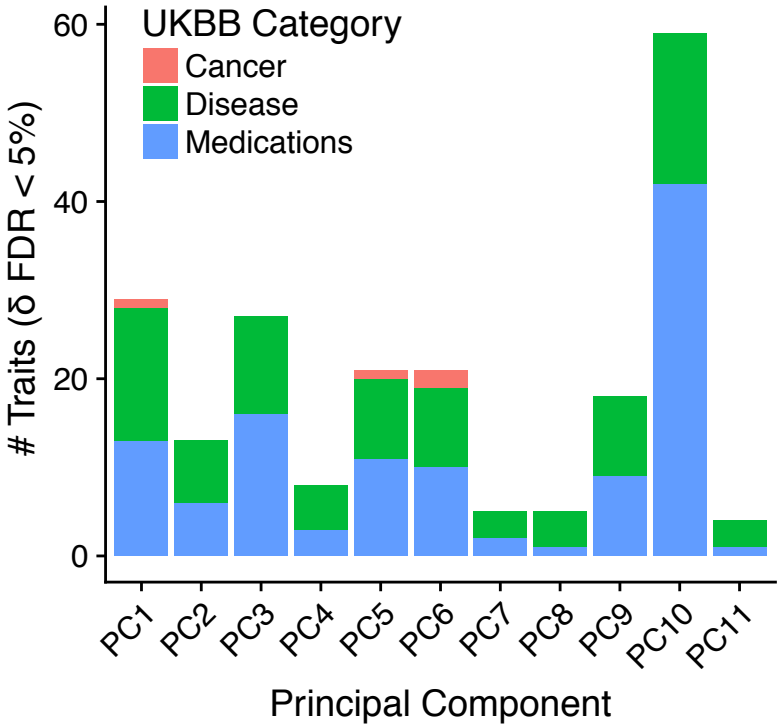


Fig. 4.12 Summary of significant UKBB SRD traits that on basis projection were found to have significant (FDR<5%) δ values. Colours represent the three categories; Cancer ($n = 28$), Disease ($n = 282$) and Medications ($n = 542$).

with diseases for which they are indications. For example Asacol and Mesalazine (Figure 4.14, negative δ) are used in the treatment of IBD, whereas Methotrexate and Thyroxine (Figure 4.14, positive δ values) products are used in the treatment of RA and Hypothyroidism respectively. This pattern repeats across most of the PC components, providing reassurance of the SRD projections. Also of interest is the significantly negative δ for basal cell carcinoma, this agrees with the finding that both non-melanoma and melanoma skin cancer rates are elevated in individuals with IBD (Long et al., 2012). Whether this risk is related to treatment with thiopurines and immunosuppressants, the genetic architecture of IBD, or combination of both remains to be elucidated.

PC3 On this component δ values for significant SRD traits on this PC are all negative, and I found no diseases with significant positive scores. Furthermore I found no nosological relationship between δs indicating that this component captures a more common genetic architecture found across all IMDs. Notable exceptions to this are MS and Asthma which have non significant

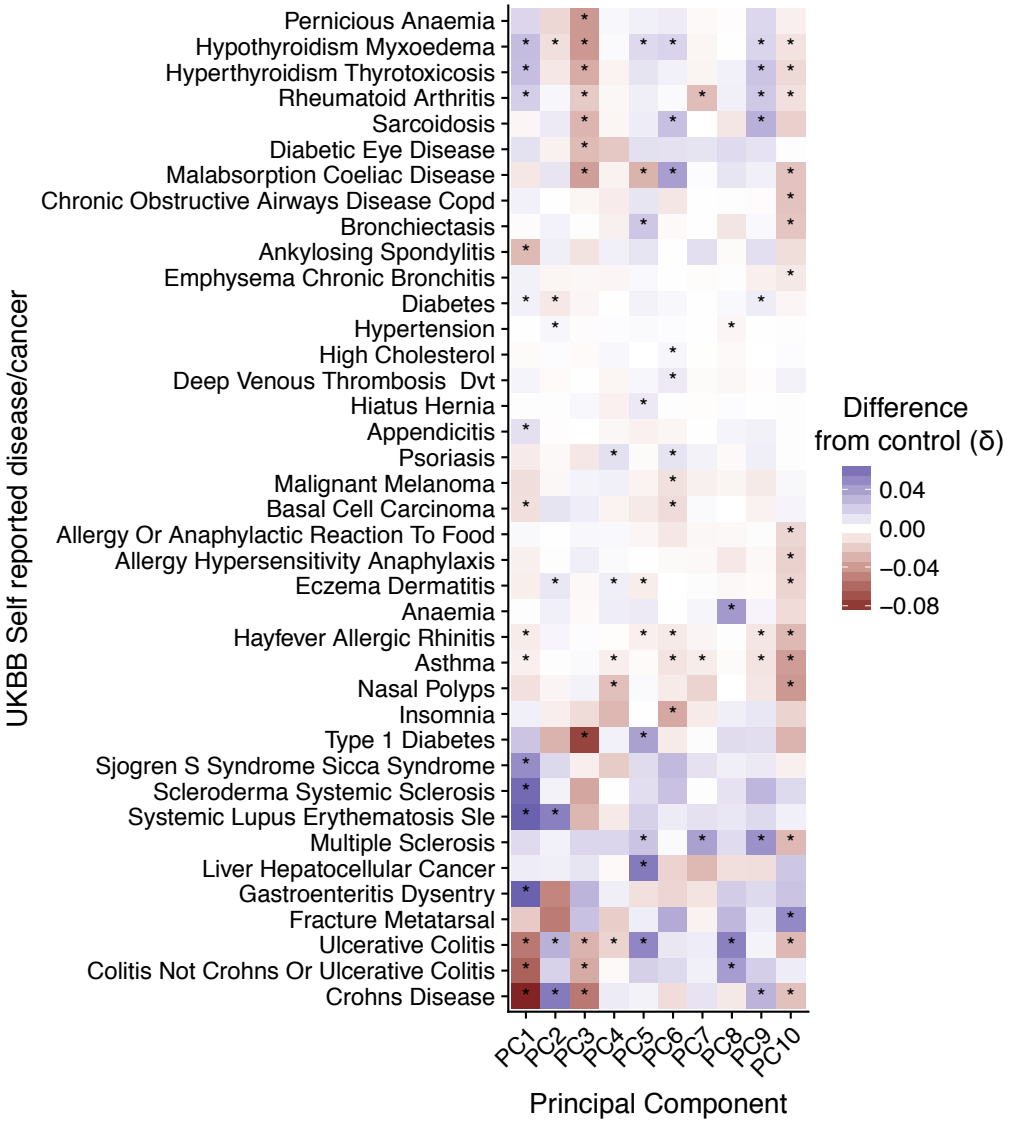


Fig. 4.13 Heatmap of significant UKBB SRD projections. PC score distance from control (δ) is indicated by colour. An asterix indicates δ is significant at FDR<5%. For clarity, PC11 is omitted as it explains a negligible amount of variance between basis traits making its interpretation challenging.

positive δ values. Interestingly, treatment response to anti-TNF therapy, a mainstay biological treatment for many immune-mediated diseases, mirrors this pattern. For example it is routinely used in the treatment of RA, which has a significantly negative δ values, whilst it is not efficatious in the treatment of asthma (Holgate et al., 2011). Indeed in multiple sclerosis which has the most positive δ values, anti-TNF therapy is a contraindication and can exacerbate disease symptoms (Cho and Feldman, 2015; Gregory et al., 2012).

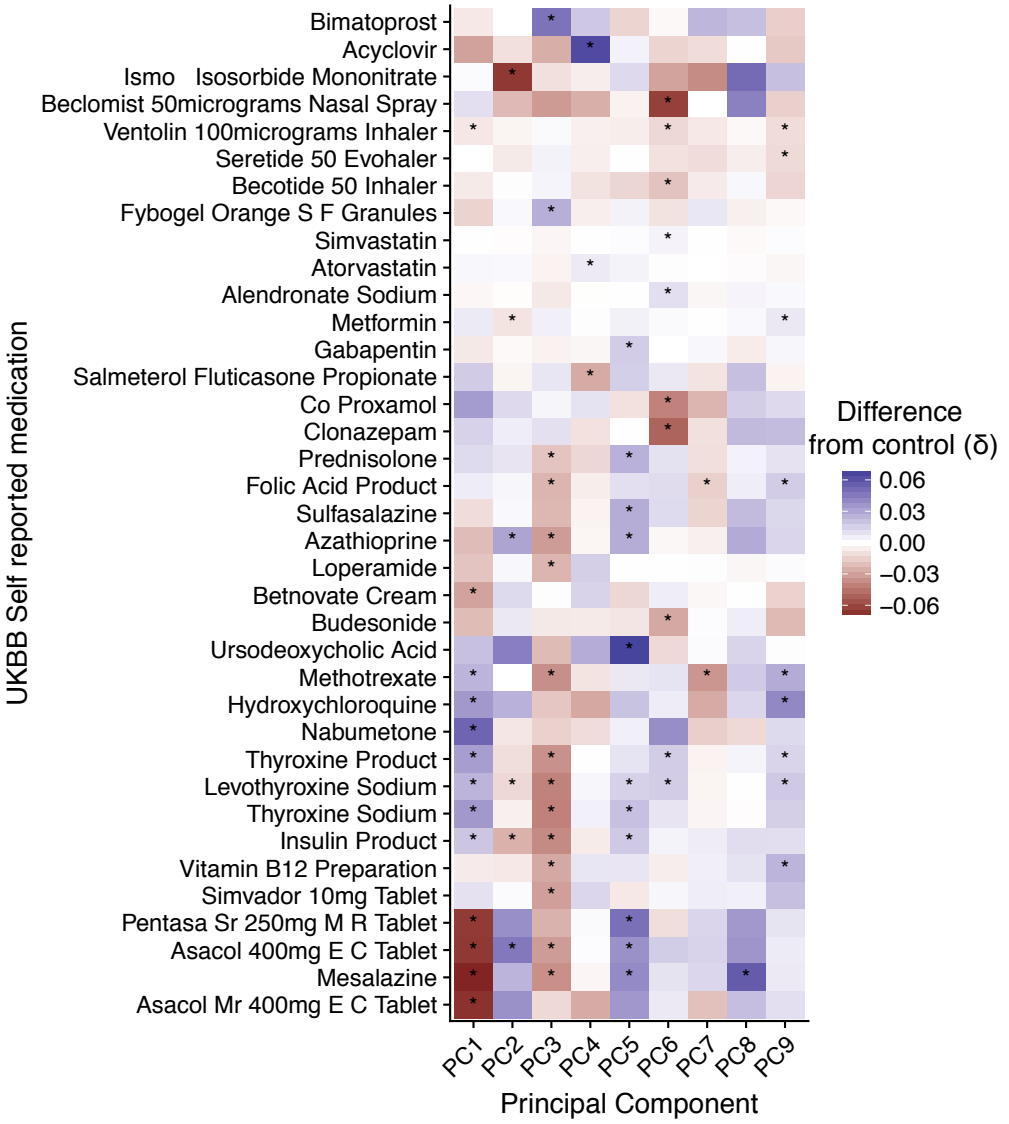


Fig. 4.14 Heatmap of significant UKBB self-reported medication. PC score distance from control (δ) is indicated by colour. An asterix indicates δ is significant at FDR<5%. A large number of asthma medications are significantly associated with PC10, and so it it is omitted to facilitate interpretation. For clarity, PC11 is omitted as it explains a negligible amount of variance between basis traits making its interpretation challenging.

PC5 This PC is defined by two basis traits, PSC and PBC that are not part of the self-reported UKBB set, having positive δ values. I found that, ursodeoxycholic acid, that is used to treat both PSC and PBC had the most positive δ values out of all projected traits. This observation supports the notion that this PC represents a Coelic disease and PBC/PSC axis. The observation that hepatocellular carcinoma (HCC), also correlates with the δ values of

PSC and PBC provides further support, as the cirrhosis endemic in both diseases is associated with increased in risk of HCC development (Lindor et al., 2009).

PC9 and PC10 Both of these PCs seem to separate asthma and allergic diseases from the other IMDs. I found that δ values between diseases and related treatments were strongly correlated. This was especially evident for PC10 where a more negative PC score captures the genetic architecture of allergic disease (e.g. asthma, Hayfever and Anaphylaxis), as well as a more general component of IMDs. I found that 34 out of 40 (85%) medications with a significantly negative δ values, were for treatments for allergy or asthma.

4.8.2 Projection of UKBB blood count GWAS

The study by Astle et al. (2016), presents a comprehensive genetic association analysis of 36 blood cell count phenotypes across 173,430 individuals of European ancestry which suggests a causal relationship between certain blood cell phenotypes and immune-mediated disease. Thus projecting summary GWAS statistics from this study onto the basis presents an opportunity to investigate whether the proposed framework is able to uncover similar relationships. Furthermore such cell type specific relationships might shed light on the causal tissue contexts underlying basis PCs.

Basis projection

I downloaded[4] GWAS summary statistics for all 36 traits, filtering each for SNPs intersecting the basis and aligned the alleles such that they were concordant with the basis. This resulted in a $36 \times 280,651$ matrix on which I applied the proposed shrinkage (Section 4.6) prior to projection onto the basis space.

In order to select significant blood trait δ values I concatenated those arising from this analysis with those derived from the UKBB self-reported binary traits (Section 4.8.1) taking forward those with a FDR < 5%, for further analysis. I focused on the 13 main blood cell indices on which Astle et al. (2016) conducted Mendelian randomisation analysis (Table 4.3), which found significant relationships between IMD risk and counts for eosinophils (Asthma and CEL), lymphocytes (Asthma, CEL and MS) and neutrophils. For eosinophil counts (EO#) they found a positive relationship between Asthma, RA, T1D and CEL

[4]http://www.bloodcellgenetics.org/

Trait	Abreviation
Lymphocyte Count	LYMPH#
Neutrophil Count	NEUT#
Eosinophil Count	EO#
Basophil Count	BASO#
Monocyte Count	MONO#
Mean Corpuscular Haemaglobin	MCH
Hematocrit (fraction of blood volume occupied by red cells)	HCT
Red Cell Distribution Width	RDW
Reticulocyte (immature red blood cells) Count	RET#
Immature Fraction of Reticulocytes	IRF
Platelet Count	PLT#
Mean Platelet Volume	MPV
Platelet Distribution Width	PDW

Table 4.3 Table of 13 main blood measurements analysed by Astle et al. (2016)

risk, whereas across the four PCs with at least one significant δ value I found conflicting evidence (Figure 4.15). For example for PC4, EO# has a negative δ, concordant with Asthma and RA but not T1D and CEL. Overall EO# PC scores show a strong concordance with Asthma but not other basis traits. For lymphocyte counts (LYMPH#) Astle et al. (2016), demonstrated a positive relationship between this trait and MS, with a lower count indicating protection from Asthma and CEL. Whilst PCs with significant δ values for this trait showed strong overlap with Asthma, CEL and MS were discordant. Finally, I found overlap for all PCs with significant δ values for neutrophil count (NEUT#) with Asthma, in agreement with Astle et al. (2016), who found a positive relationship between Asthma risk and NEUT#.

Overall I found these results difficult to interpret outside of the established association between (EO#) and asthma (Bousquet et al., 1990). The disappointing overlap with projected blood traits and the causal overlaps suggested by Astle et al. (2016), is perhaps, unsurprising as there are many technical differences between the two approaches. For example, there is a large difference in both the number and ascertainment method as to which SNPs were included in both analysis method. Whilst the basis is restricted to considering the subset of SNPs present across all studies, Astle et al. (2016) considered IMD SNPs that were *a priori* in high LD with one or more sentinel SNPs associated with one or more of the 13 main blood traits (MAF>0.01%). Such a bias in terms of both SNPs considered and how they were ascertained is likely to lead to differing results regardless of the analytical method used.

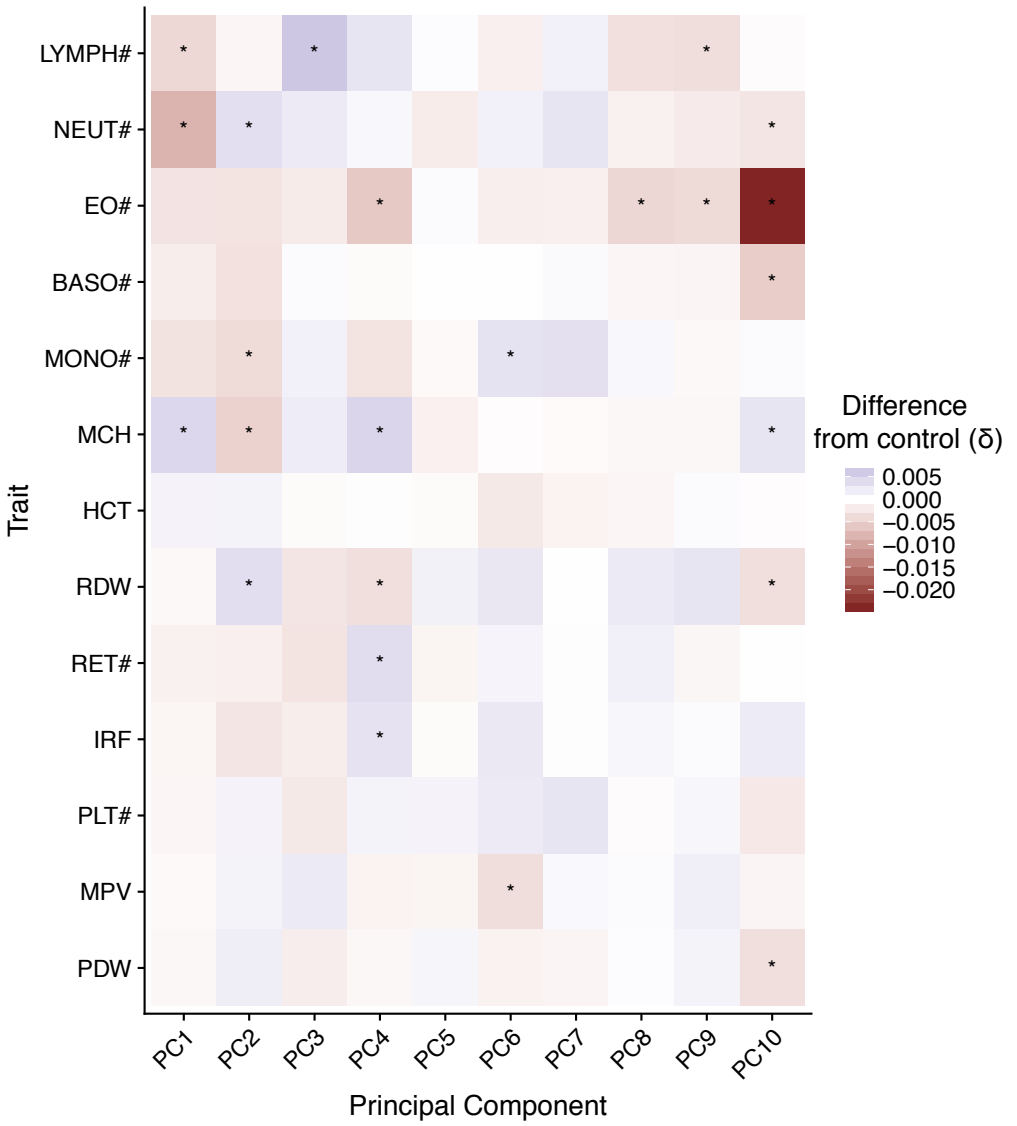


Fig. 4.15 Results of projection of 13 main blood count traits from Astle et al. (2016), see Table 4.3 for trait abbreviations. PC score distance from control (δ) is indicated by colour. An asterix indicates δ is significant at FDR<5%.

4.8.3 Projection of whole blood eQTL data

One approach to characterising the biology underlying basis components is the integration of genetic data influencing gene expression in a relevant tissue type as this has the potential to link the genetic regulation of the expression of specific genes to a component. Recently, Võsa et al. (2018) performed a meta-analysis of whole blood eQTL data on a total of 31,684 healthy individuals of European ancestry. I sought to project this dataset onto the basis in order provide a richer characterisation of basis components.

I downloaded summary statistics from http://www.eqtlgen.org/cis-eqtls.html for cis eQTLs, from testing for association between a given gene's expression (across 19,960 genes expressed in blood) and each SNP within a 1Mb region from the gene centre. To this I added summary statics downloaded from http://www.eqtlgen.org/trans-eqtls.html for what Võsa et al. (2018) term trans-eQTLs, that result from association testing between a given gene's expression and each SNP from a curated list of 10,317 trait-associated SNPs outside of a 5Mb region from the gene centre.

For each gene, I filtered summary statistics to obtain only those for SNPs included in the basis. Given that the summary statistics available are predominantly centred on a given gene this leads to large amounts of missing data, where an eQTL summary statistic is not present for most basis SNPs. Over 19,942 genes for chromosomes included in the basis, the mean number of variants with summary statistics was 1,804 (IQR 1,742-1,861). I set effect sizes for variants where summary statistics were not available to zero under the reasonable assumption that untested or unreported variants were likely to have negligible effects sizes on target gene expression. After aligning effect alleles and performing Bayesian shrinkage I projected all genes onto the basis, resulting in a total of 11 component scores for each of 19,942 genes for which at least one basis SNP was available.

I generated Z scores using empirical mean and variance estimates across all gene projections for a given PC such that

$$Z_{kg} = \frac{P_{kg} - \bar{P}_k}{\sqrt{\mathrm{Var}(P_k)}}, \tag{4.24}$$

where P_{kg}, is the PC score for the k^{th} component associated with projecting a gene, g onto the basis (Figure 4.16).

I devised a gene set enrichment strategy in order to assess whether components were associated with specific biological gene-sets through eQTLgen gene projections. To do this I downloaded Reactome (Fabregat et al., 2016), KEGG (Kanehisa and Goto, 2000) and HALLMARK gene-sets in gmt format from http://software.broadinstitute.org/gsea/msigdb (Liberzon et al., 2015). I created a gene 'universe' by selecting the unique set of identifiers for protein coding genes for which PC scores were available across all components. I used this to filter pathways such that genes without a PC score were removed, this resulted in pathways containing very few genes, to counter this I only took forward those with more than 10 genes for downstream analysis. Given that for a component, genes can have both positive and negative PC scores, using the raw PC scores for

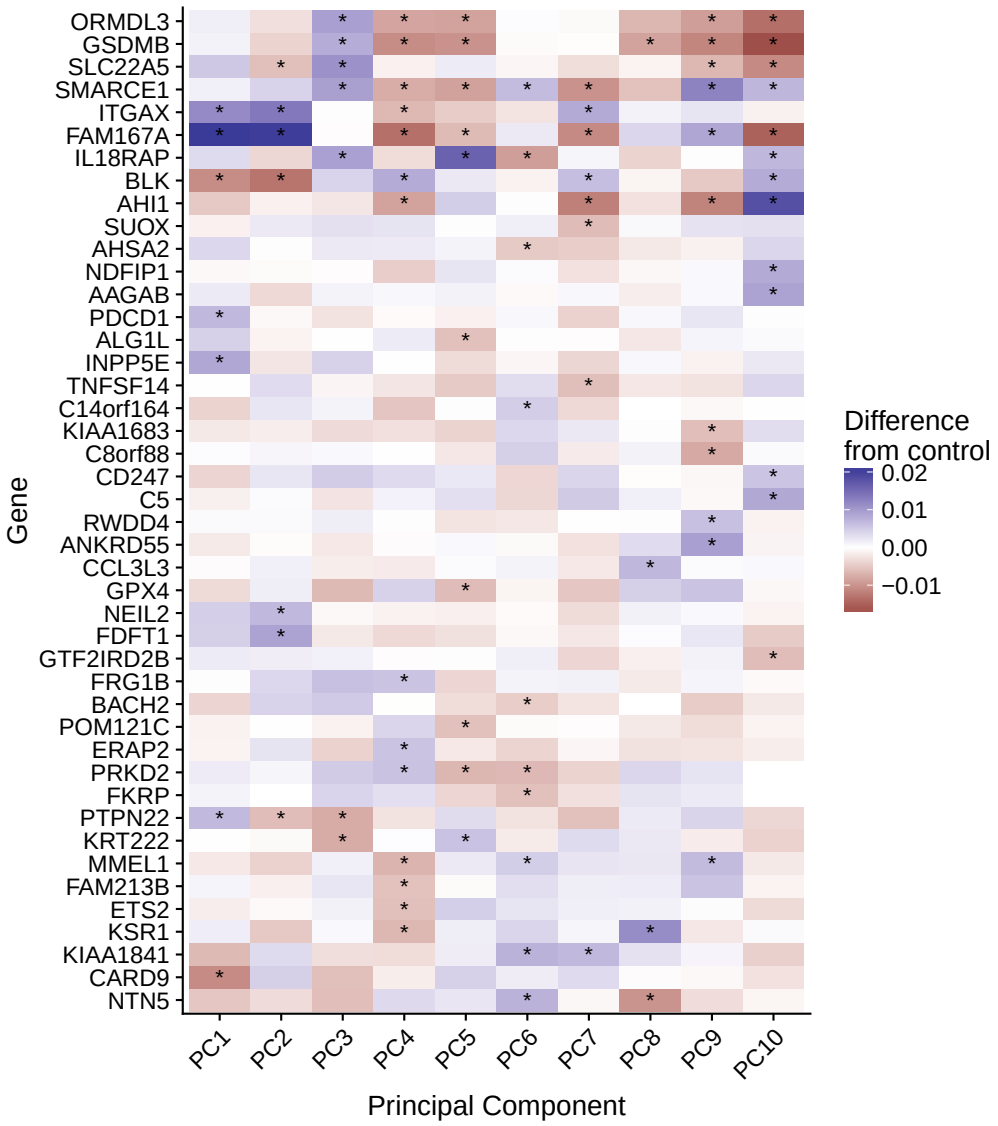


Fig. 4.16 Heatmap of significant gene projections from Võsa et al. (2018). Colour indicates difference from control PC score, '*' denotes significant at FDR < 5%.

enrichment might be misleading as these effects might be antagonistic, resulting in no net enrichment. A more robust approach is to compare the variance of PC scores between genes within a set and their complement. To do this I used the `var.test` function in R which implements an F-test on the ratio of variances of PC scores of genes in the two groups. I used this approach to assess each of the filtered gene-sets in order to identify those with evidence of large PC score variance.

As might be expected I found that gene-sets with relevance to both adaptive and innate immunity showed evidence of enrichment (Figure 4.17) across all basis

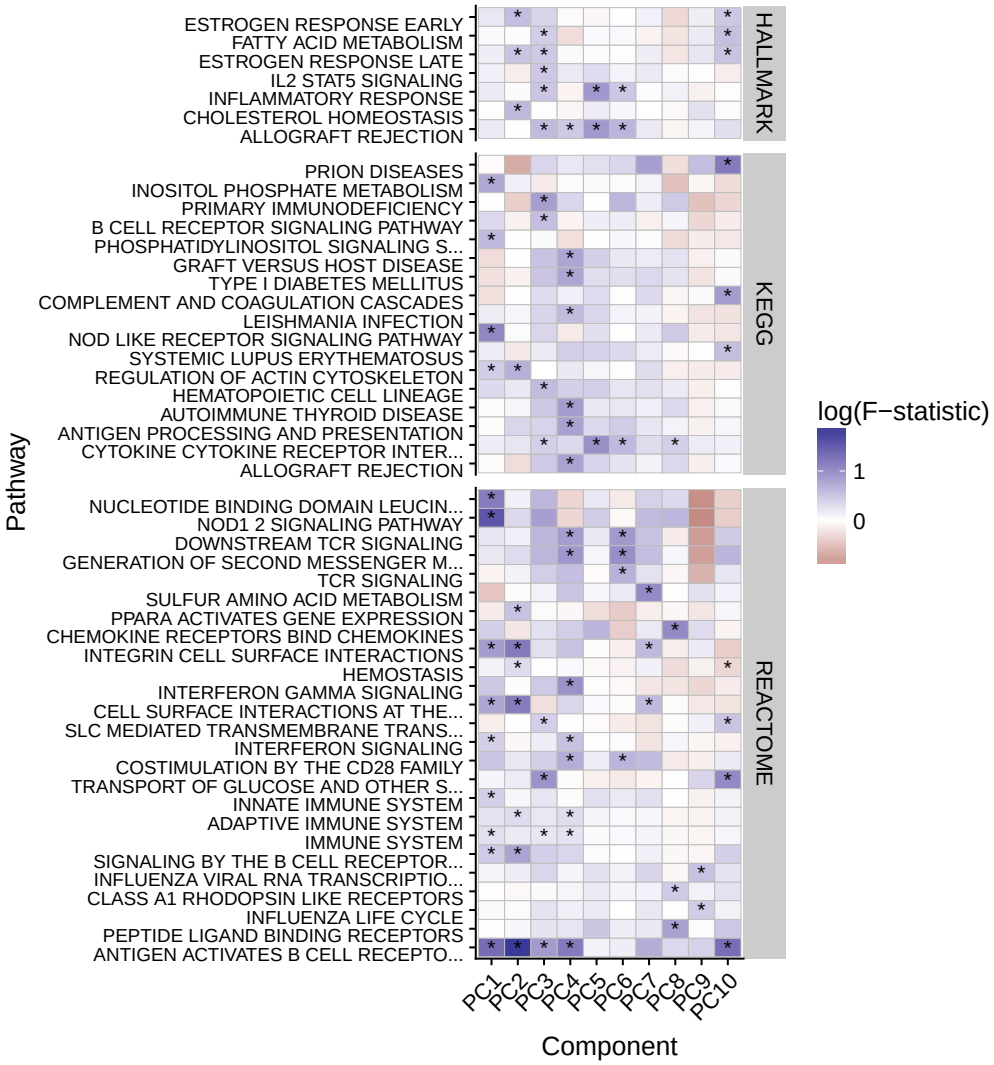


Fig. 4.17 Heatmap of variance enriched pathways from projection of variance enriched pathways from projection of Võsa et al. (2018). Colour indicates $\log(F\text{-statistic})$ from an F-test of equality of variances , '*' denotes significant at FDR < 1%.

components. Specifically, PC1 was characterised by both inflammatory pathways (e..g NOD like receptor signalling and integrin cell surface interactions (de Lange et al., 2017)) and those involved in adaptive immunity (e.g. b-cell receptor signalling), supporting its role in describing an autoinflammatory/autoimmune axis. Pathway associations with other components were less clearly defined, with more generalised roles for metabolism (e.g. fatty acid metabolism), cytokine signalling (IL2/STAT5 signalling) and T and B cell signalling. When projecting on UKBB SRD (Section 4.8.1) I observed an association between Asthma and the medication used in its treatment on PC10, the association of the coagulation/complement sys-

tem provides further support for this as the importance of this biological pathway with respect to Asthma has previously been reported (de Boer et al., 2012).

4.9 Using the basis to characterise JIA

In the previous sections I proposed a framework for creating a basis that facilitates the comparison of the genetic architecture of clinically and/or phenotypically related disease by creating a low dimensional summary using publicly available GWAS summary statistics. Such a representation appears to mirror disease nosology, with for example PC1 discriminating between auto-inflammatory and more autoimmune diseases (Section 4.6.2). In an extension, by projecting new datasets into basis space, demonstrated using UKBB SRDs, I was able to show that basis was able to cluster these on the basis of shared genetic architectures (Figure 4.10).

In this section, I expand on this application, describing how it can be used to better understand the genetic difference underpinning a clinically heterogenous disease, using juvenile idiopathic arthritis (JIA) as a motivating example.

4.9.1 JIA disease subtypes

JIA is the most common cause of chronic childhood rheumatic disease, with heterogeneous arthritides occuring prior to the age of 16 years and persisting for at least 6 weeks (Ravelli and Martini, 2007). This heterogenity has led to the International League of Associations for Rheumatology (ILAR) criteria, which divides JIA into subtypes on the basis of clinical features (Petty et al., 2004) and comprises seven main disease subtypes (Table 4.4).

In previous work Hinks et al. (2017), investigated the HLA region in order to suggest genetic associations that were common or distinct between subtypes. This analysis was by design restricted to examining disease overlap using SNPs within the HLA region, but nevertheless suggested an overlap between polyoligoarthritis (PO), extended oligoarthritis (EO) and rheumatoid factor negative (RF-), the most common disease subtypes. In addition to this Ombrello et al. (2017), using a cohort of 770 children, found no evidence for a shared genetic architecture between systemic JIA (Sys) and the other JIA disease subtypes, supporting the observation that Sys is a distinct disease. However, the method employed to examine overlap was limited to examining either targeted regions (Prahalad et al., 2013) or those covered by ImmunoChip genotyping platform (Hinks et al., 2013) and so was not a full genome-wide analysis.

Subtype	Abbreviation	Case#	Phenotype
Psoriatic	PsA	150	Psoriasis/ dactylitis/ onycholysis
Enthesitis-related	ERA	185	Ankylosing spondilytis/ Sacroiliac joint tenderness/ inflammatory lumbosacral pain/ HLA-B27
Polyoligo RF+	RF+	199	Five or more joints and Rheumatoid factor positive
Systemic	Sys	283	Daily fever
Extended oligo	EO	394	Five or more joints after 6 months
Polyoligo RF-	RF-	573	Five or more joints and Rheumatoid factor negative
Persistant oligo	PO	650	Four or fewer joints

Table 4.4 JIA disease subtype cohort description. Disease subtypes and their clinical phenotypes and are derived from (Hinks et al., 2017). Case# indicates the number of individuals with a particular ILAR classification of JIA disease subtype for which genotype data was available for analysis in this thesis.

4.9.2 JIA subtype GWAS analysis

I obtained quality controlled genotype, data from collaborators at The University of Manchester, for samples for all seven ILAR coded subtypes of JIA (Table 4.4) and 5,181 shared controls. I had no input into the collection, recruitment or processing of genotyping calls. I was supplied with a matrix of PC scores derived from PCA of the genotype matrix, performed by the Manchester group in order to estimate and thus control for subtle ancestry specific effects that might confound subsequent analysis.

I used this data to perform a case/control GWAS of JIA subtypes limited to the variants contained in the basis using the same 5,181 controls for analysis across all subtypes. Briefly, I used the R package snpStats (Clayton and Leung, 2007) to fit a logistic model, for each of the subtypes using sex and as recommended by the Manchester group the first three PCs obtained from PCA of the genotype matrix as covariates.

4.9.3 Projection of JIA subtype GWAS

The GWAS performed in the previous section resulted in a set of summary statistics for each of the 7 subtypes. After aligning effect alleles with the basis, I applied the

weights as previously described to $\hat{\beta}$, and projected each subtype onto the basis. This projection resulted in a 7×11 matrix of projected PC scores (Figure 4.18). The most striking difference that I observed was on PC3 where scores for all subtypes except ERA and systemic JIA were significantly different to the 'control' basis trait.

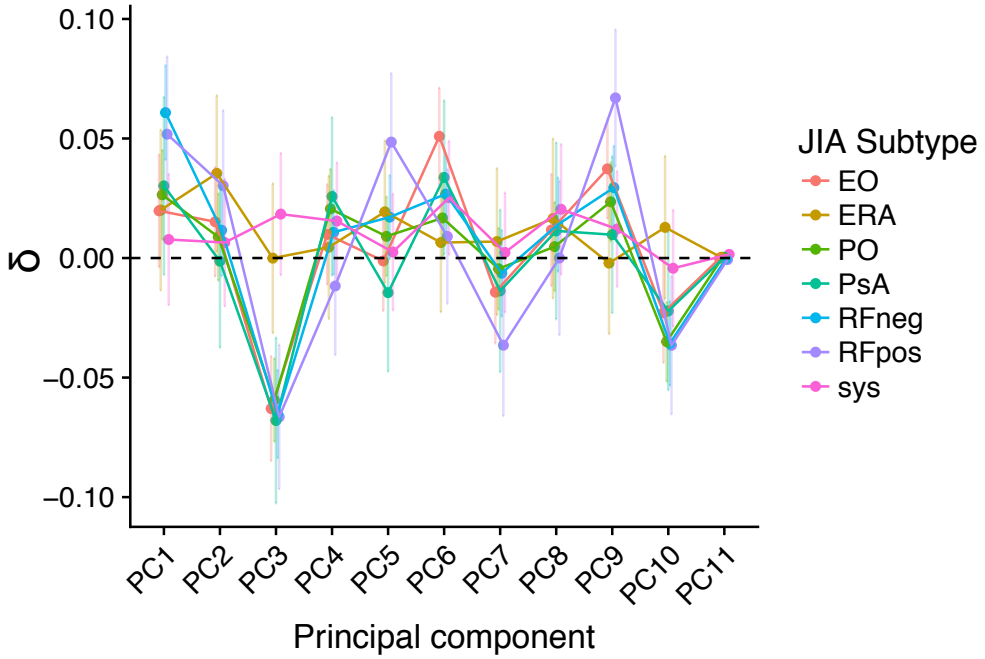


Fig. 4.18 Control centred projected PC scores across 7 JIA subtypes. The y axis represents the difference between the 'control' PC score and the projected PC scores for a given trait (δ). For clarity faded lines indicate 95% confidence intervals for the projection, calculated using the method proposed in section 4.7

PC3 discriminates ERA and Systemic JIA from other subtypes

In section 4.7 I introduced a method for analytically computing the variance of a PC score for a projected trait. This showed that whilst PC3 δ values for RFneg, PO, EO, RFpos and PsA were significant even after employing Bonferroni multiple testing correction, both ERA and Sys were not (Table 4.5).

Comparison with HLA focused Hinks et al. (2017) subtype analysis

Whist I observed striking differences at PC3 it is important to consider other PCs in order to get an overall view of the similarities and difference between the genetic architecture of JIA disease subtypes. To do this I applied the hierarchical

Trait	Cases	Z	p-value	$p_{\text{adj.}}$
RFneg	573	-7.0	3.4×10^{-12}	2.6×10^{-10}
PO	650	-6.7	2.1×10^{-11}	1.6×10^{-9}
EO	394	-5.7	1.5×10^{-8}	1.2×10^{-6}
RFpos	199	-4.3	1.6×10^{-5}	1.2×10^{-3}
PsA	150	-3.9	1.2×10^{-4}	9.0×10^{-3}
ERA	185	0.0	≈ 1	≈ 1
Sys	283	1.4	0.16	≈ 1

Table 4.5 Table of significance compared to control traits for JIA subtypes on PC3. The Adj. p value column represents the Bonferroni corrected values across all 11 PCs and disease subtypes.

clustering approach previously described (Section 4.5.2) to the matrix of PC scores for the 7 projected JIA subtypes. This analysis (Figure 4.19a) suggested three clusters, the largest of which contained RFneg, EO, PO and PSA with a further cluster containing ERA and systemic subtypes. Interestingly RFpos appeared to exhibit only limited overlap with other disease subtypes.

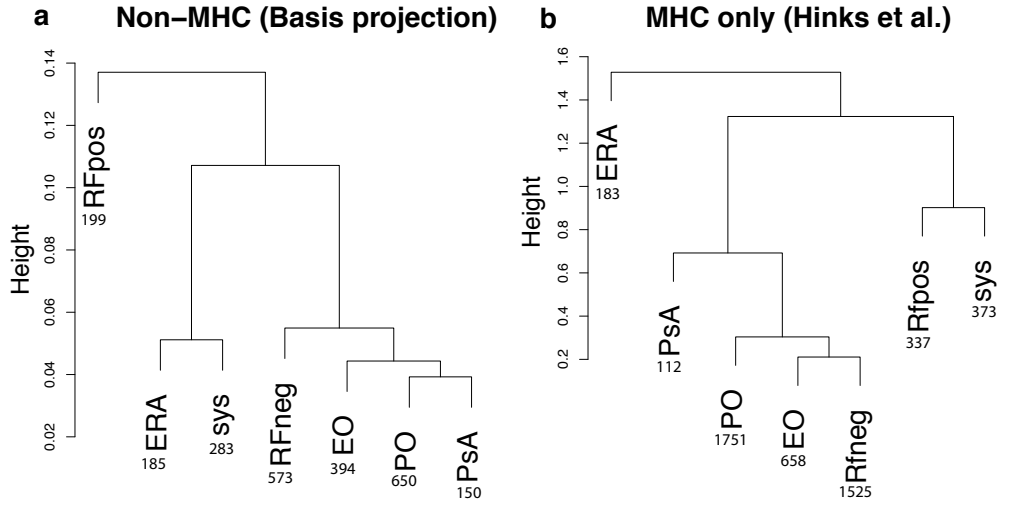


Fig. 4.19 **a**) clustering using basis Euclidean distance of PC scores. **b**) clustering using correlation data taken from (Hinks et al., 2017). Numbers under each label indicate the number of samples analysed.

As previously mentioned Hinks et al. (2017), used genetic correlation, conducted using (GATK), across the HLA region in order to characterise the pairwise genetic overlap between disease subtypes. For comparison with the approach detailed here I applied the hierarchical clustering approach to the matrix of genetic correlations presented in Hinks et al. (2017). It should be noted that whilst there is significant overlap between the individuals included between the two studies

they are not identical. Interestingly, I found that my analysis identified a similar cluster containing RFneg, EO, PO and PSA albeit with a slightly modified fine grained structure, even though the basis approach explicitly excludes the HLA region from consideration. In Hinks et al. (2017), the clustering of PO, EO and RFneg is mostly likely explained by the shared susceptibility allele at *HLA-DRB1* amino acid glycine 13. Conversely whilst a histidine at *HLA-DRB1* amino acid 13 protects from these subtypes it increases risk of developing the RFpos subtype, perhaps explaining why RFpos clusters separately. Hinks et al. (2017) suggest that PsA might be susceptible to misclassification, and therefore might comprise a mixture of PO, EO and RFneg subtypes. The similarities between the results obtained by the two approaches indicate that at least for PsA, EO, PO, RFneg, there is a common genetic architecture that exists outside of the major HLA-DRB1 locus and that non-HLA loci are at least in part responsible for a shared genetic architecture between these traits.

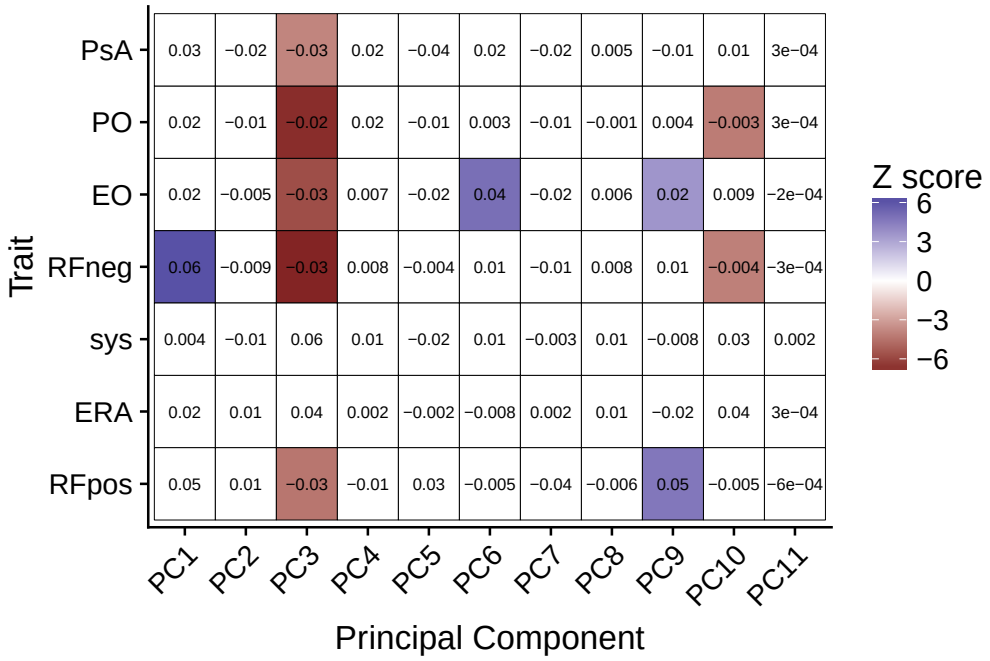


Fig. 4.20 Heatmap of JIA subtype projections, cells are labeled with δ values and coloured by Z score if they exceed Bonferonni significance.

This leads to the question as to whether their are additional components, exclusive of PC3, that have significantly different PC scores for JIA subtypes (Figure 4.20). Finding significance is heavily dependent on sample size (Equation 4.20), as such I found that EO and RFneg, subtypes with the greatest number of cases, had the most components with significant δ values, after Bonferroni

correction. Interestingly, both Sys and ERA had no significant δ values for any component, whilst some of this is no doubt due to sample size, I note that the smallest disease subset, PsA is significantly associated with PC3. Such an observation is also supported by the analysis of Ombrello et al. (2017), who found that using polygenic risk scores built from combinations of the most common subtypes (PO,RFneg and RFpos), had no predictive power when applied to a cohort of systemic JIA patients.

In summary, the projections of JIA subtypes into basis space that I observed appear to have some support from alternative studies (Hinks et al., 2017; Ombrello et al., 2017). However, the results do raise the important question as to what, biologically or clinically, the principal components represent. Furthermore, when comparing between subtypes (rather than with control) care needs to be taken due to the fact that input summary statistics are computed using a shared pool of control subjects, which I discuss further in Section 4.9.5.

4.9.4 Annotating PCs related to JIA

The characterisation of components as described previously (Section 4.8) is valuable as it can be used to gain knowledge about how individual components relate to JIA disease subtypes. For PC1 all JIA subtypes have positive δ values indicating a tendency towards autoimmunity however, only PO, RFneg and RFpos show significance (FDR<0.05). The ERA subtype has clinical and genetic (through the presence of HLA-B27) similarities with adult-onset ankylosing spondylitis (Colbert, 2010). However, I observed significantly different PC1 scores (t-test $p = 0.01$) between UKBB SRD ankylosing spondylitis ($N = 1,058$) and ERA with a tendency of ERA towards an autoimmune phenotype. One explanation for this might be a tendency for the misclassification of other forms of JIA to the ERA subtype based on a reliance on the the presence of HLA-B27 (MAF approx 7%). Unfortunately summary statistics for a large AS GWAS outside of UKBB, such as that presented in International Genetics of Ankylosing Spondylitis Consortium (IGAS) et al. (2013) are not publicly available, precluding validation of this finding.

Returning to PC3 which separates Sys and ERA from other disease subtypes, I previously described (Section 4.8.1) how this component also separated self-reported MS and asthma from other IMDs and a possible link with anti-TNF therapy. However a more complicated picture emerges when JIA is taken into account, as whilst there is limited evidence for the efficacy of anti-TNF treatments

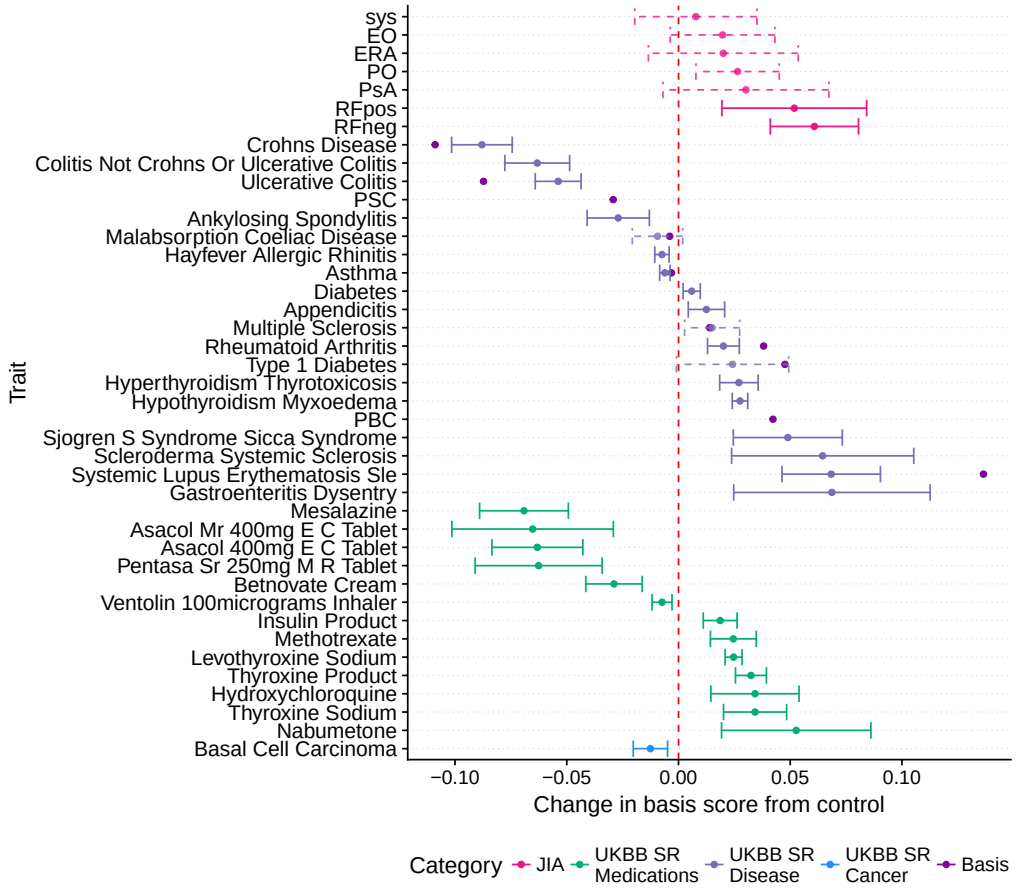


Fig. 4.21 Forest plot showing the context of JIA subtype projections for PC1. Coloured points indicate the difference in trait scores with synthetic control, marked with a red dotted line. Error bars indicate 95% confidence intervals. For clarity basis trait PC scores (purple) have been merged with overlapping UKBB traits. Whilst non-basis UKBB traits have been filtered to show only those that are significantly different from control at FDR < 5%, other traits are included regardless, these are shown with dashed error bars.

in treating systemic JIA (Russo and Katsicas, 2009), beneficial outcomes have been observed in children with ERA (Gmuca et al., 2017; Weiss et al., 2018).

I did not observe a specific overlap between PsA and self-reported Psoriasis for components with significant δs, perhaps due to the misclassification within the PsA subtypes suggested by Hinks et al. (2017). Forest plots for all PCs such as the example in Figure 4.22 are available in Appendix C.2.

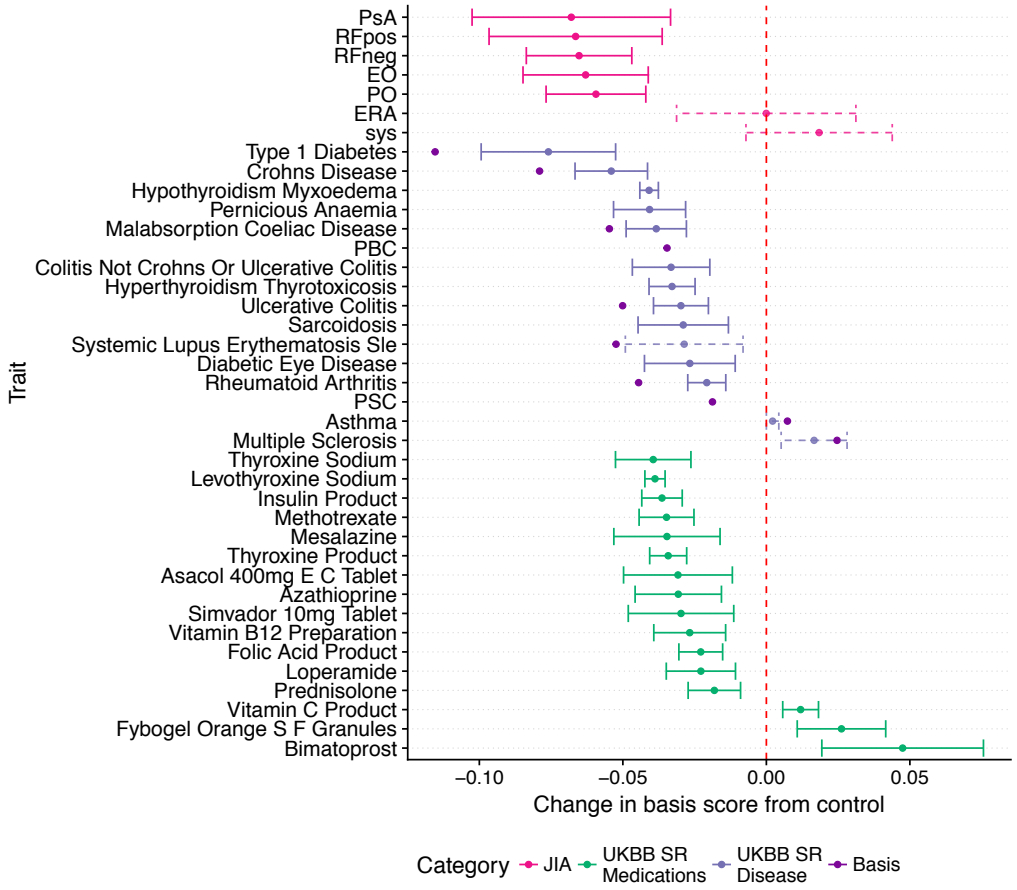


Fig. 4.22 Forest plot showing the context of JIA subtype projections for PC3. For further details see Figure 4.21 legend.

4.9.5 Comparing JIA subtypes PC scores in the presence of shared controls

Up to this point I have focused on a statistical method for comparing a projected disease PC score with a synthetic basis control trait, where no effect is assumed at each basis SNP. A natural approach to testing the significance of principal component scores between sets of traits is the t-statistic which involves computing the the pooled mean and variance of the sets of PC scores to be compared. This is complicated when comparisons involve study designs that utilise a common pool of samples, where pooled variance will be underestimated if independence is assumed. Lin and Sullivan (2009) show that the correlation between study effect estimates with shared samples can be estimated as follows:

$$\mathrm{Corr}(\hat{\beta}_k, \hat{\beta}_l) = \left(n_{kl0} \sqrt{\frac{n_{k1} n_{l1}}{n_{k0} n_{k0}}} + n_{kl1} \sqrt{\frac{n_{k0} n_{l0}}{n_{k1} n_{k1}}} \right) \Big/ \sqrt{n_k n_l} = \rho_{kl}, \qquad (4.25)$$

where n_{k1}, n_{k0}, and n_k (or n_{l1}, n_{l0}, and n_l) are, respectively, the number of cases, the number of controls, and the total number of subjects in the k^{th} (or l^{th}) study and n_{kl0} and $n_{kl1}$are the total number of controls and cases that overlap. Let $X = \{x_1, \ldots, x_{n_x-1}, x_{n_x}\}$ and $Y = \{y_1, \ldots, y_{n_y-1}, y_{n_y}\}$, be vectors of the PC scores for two groups of traits for a given PC. The variance of the difference between the means of the two groups is

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{\sum_i^{n_x} \mathrm{Var}(x_i)}{n_x^2} + \frac{\sum_j^{n_y} \mathrm{Var}(y_j)}{n_y^2} - \frac{2}{n_x n_y} \sum_{i,j} \mathrm{Cov}(x_i, y_j), \qquad (4.26)$$

where, by (Equation 4.25), $\mathrm{Cov}(x_i, y_i) = \rho_{kl} \sqrt{\mathrm{Var}(x_i)\mathrm{Var}(y_i)}$. This allows the computation of a t-statistic $t = \frac{\bar{X}-\bar{Y}}{\sqrt{\mathrm{Var}(\bar{X}-\bar{Y})}}$ that is appropriately adjusted for shared cases or controls.

I applied this method to examine, for PC3 the difference in projected PC scores between two sets of JIA subtypes, with one set containing ERA and sys, and the other containing all other subtypes. As expected I found this to be particularly significant ($p = 1.66 \times 10^{-9}$), even taking into account shared controls. For comparison, the application of a pooled variance approach without taking into account shared controls was inflated ($p = 4.58 \times 10^{-10}$), and whilst this makes no difference to the conclusion for this comparison, for more subtle comparisons sample sharing will affect type 1 error.

4.10 Projecting individual genotypes onto the basis

Up to this point I have considered the projection of effect size estimates only, however extending the method to allow as input individual genotypic data allows other applications and datasets to be applied to the method. This is because the PC scores for each PC component derived from such an application apply to an individual rather than a specific trait. For example if we project on the genotype data for a set of individuals for whom we also have gene expression data we might look for evidence for the correlation between individual level PC scores and gene expression.

4.10.1 Computation of posterior odds ratios

The projection of individual level genotype data into the basis space requires the conversion of specific genotypes onto the odds ratio scale. To do this I used a Bayesian framework developed by Chris Wallace inspired by the work of Aitkin and Chadwick (2003). This framework assumes that, at any SNP, we wish to estimate the individual odds ratio as determined by the allele frequency in controls, and a single case. In subsequent sections I use $\text{Bin}(n,p)$ to represent a binomial distribution of n trials with a probability of success p and $\text{Beta}(\alpha,\beta)$ to represent a Beta distribution with shape parameters α and β.

I assume that genotypes in controls follow a $\text{Bin}(2,f_0)$ distribution, where $f_0 \sim \text{Beta}(\alpha_0,\beta_0)$. I also assume a large number of controls, such that the parameters (α_0,β_0) may be estimated by maximum likelihood (equal to maximum *a posteriori* estimates assuming a flat prior on f_0). Assuming f_0 has been estimated as $\hat{f}_0$ the observed frequency of a specific allele in n_0 control subjects, and Hardy-Weinberg Equilibrium, I equate the estimated and theoretical expectation and variance of f_0

$$E(f_0) = \hat{f}_0 = \frac{\alpha_0}{\alpha_0+\beta_0}$$
$$V(f_0) = \frac{\hat{f}_0(1-\hat{f}_0)}{2n_0} = \frac{\alpha_0\beta_0}{(\alpha_0+\beta_0)^2(\alpha_0+\beta_0+1)}$$

and solve for

$$\alpha_0 = (2n_0-1)\hat{f}_0$$
$$\beta_0 = (2n_0-1)(1-\hat{f}_0)$$

If f_1 is the allele frequency in cases, with prior distribution $\text{Beta}(\alpha_1,\beta_1)$, and we observe a single case $G \sim \text{Bin}(2,f_1)$, then the posterior distribution of $f_1|G \sim \text{Beta}(\alpha_1+G,\beta_1+2-G)$.

To estimate (α_1,β_1), I set prior limits on the log odds ratio, Ω. First I assume that

$$\begin{aligned} E(\log\Omega) &= E\left(\log(f_0)+\log(1-f_1)-\log(1-f_0)-\log(f_1)\right) \\ &= \psi(\alpha_0)-\psi(\beta_0)-\psi(\alpha_1)+\psi(\beta_1) \\ &= 0 \end{aligned} \tag{4.27}$$

where we make use of the relations that if $f_0 \sim \text{Beta}(\alpha_0, \beta_0)$, then

$$1 - f_0 \sim \text{Beta}(\beta_0, \alpha_0), \tag{4.28}$$

and

$$E(\log(f_0)) = \psi(\alpha_0) - \psi(\alpha_0 + \beta_0) \tag{4.29}$$

where $\psi()$ denotes the digamma function. Second, we assume that

$$\begin{aligned} \Pr(|\log \Omega| > \log 2) &= \int_0^1 \Pr(l(f_0) < f_1 < u(f_0)) f(f_0) df_0 \\ &= 1 - \varepsilon \end{aligned} \tag{4.30}$$

where $l(f_0)$ and $u(f_0)$ are the lower and upper limits on f_1 corresponding to $\Omega = \frac{1}{2}$ and 2, respectively, and ε is the small probability that f_1 lies outside these limits, i.e. we assume large odds ratios are *a priori* unlikely. I used the R function `uniroot` to optimise conditions (4.27)– (4.30) to estimate (α_1, β_1) and therefore generate a function to convert individual genotypes to the odds ratio scale, given only allele frequencies in control subjects.

4.10.2 Effect of parmameters on posterior log(OR) estimates

The Bayesian framework depends upon two parameters, control sample size and the prior probability that an odds ratio exceeds a target log odds ratio, Ω, of 2 (See previous section). To analyse the effect of altering these parameters I first examined the behaviour of the framework over a range of control sample sizes (Figure 4.23a). Reflecting the larger variance of odds ratios at small control sample sizes and minor allele frequencies, I observed a small attenuation in posterior odds ratios with decreasing control sample size, although across the broad range of sample sizes this made relatively little difference. More generally the selection of a lower control sample size leads to more conservative estimates of posterior odds ratio estimates. I next examined how the prior probability for odds ratios to exceed 2 might affect posterior odds ratio estimations (Figure 4.23b). As $P(|\log(\text{OR})| > \log(2))$ decreases I observed that the relationship between allele frequency and posterior odds ratios depends much more strongly on the prior than on sample size.

Reflecting on these observations I decided to take forward the empirically obtained curves from using a sample size of 2,500 and $P(|\log(\text{OR})| > \log(2) =$

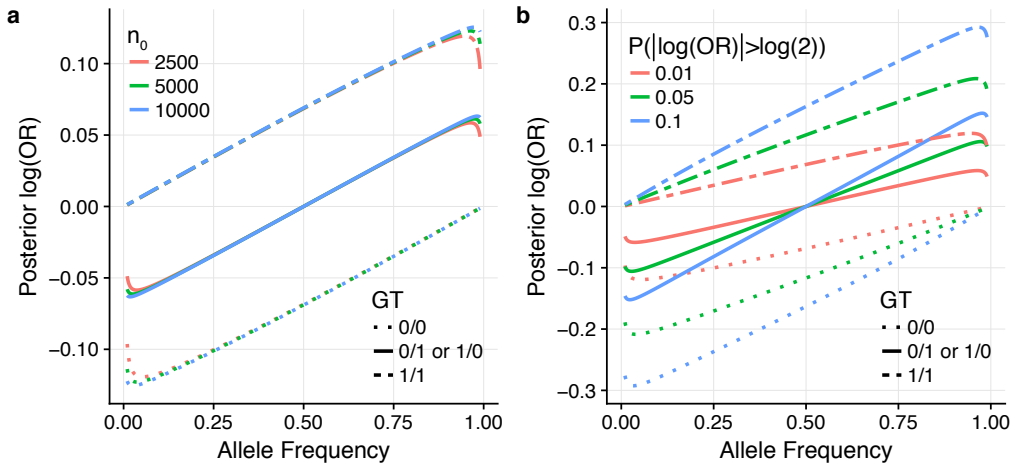


Fig. 4.23 Effect of parameters on posterior log(OR) values computed through the proposed Bayesian framework. Effect of **a)** Control sample size (n_0) and **b)** $P(|\log(\text{OR})| > \log(2))$, with sample size fixed at 2,500.

0.01 on the basis that this would generate conservative estimates for posterior odds ratios for downstream analysis.

4.10.3 Projection of JIA disease subtype genotype data into basis space

I applied the proposed method to raw genotype data for the JIA case cohort previously described (Table 4.4). This involved aligning genotypes such that reference and alternate alleles matched those used to construct the basis. Then on a per individual basis I computed an estimate of the logarithm of the posterior odds ratio, $\hat{\beta}$, for each genotype across each variant included in the basis. This required two variables, the raw genotype for the individual at that variant and the allele frequency of the alternate allele, for the latter I used allele frequency estimates from the UK10K reference cohort for consistency with the basis. Such a computation involves using the curves previously obtained with the selected parameters (Section 4.10.2) such that given an observed genotype and control allele frequency, the required posterior log odds ratio can be simply looked up. After applying the standard basis shrinkage to $\hat{\beta}$ across all basis SNPs I projected each individual into basis space.

One approach to the assessment of the PC scores obtained from the proposed genotype level method is to investigate how they compare to those obtained from projecting $\hat{\beta}$ estimates from case/control GWAS (Section 4.9.3). In order to prepare a comparable summary I took the mean PC score, across individuals

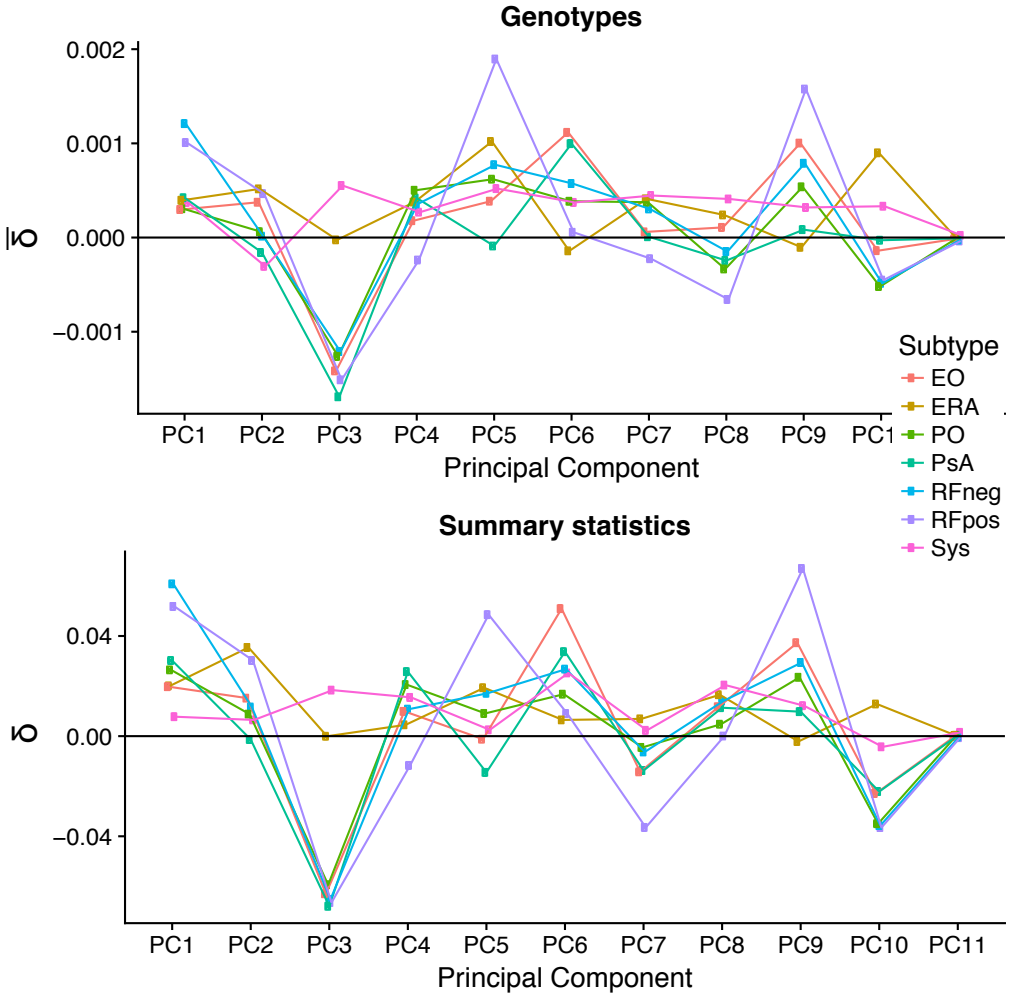


Fig. 4.24 Comparison of JIA subtype PC δ values derived from genotype (top panel) and summary data approaches (bottom panel).

for each JIA subtype and plotted them along side the summary method results ((Figure 4.24) Whilst the scaling of PC scores for the genotype method was much reduced compared to the summary statistic method, the overall patterns across disease subtypes was conserved (Figure 4.24) offering support for the proposed framework. I found that a t-test of individual PC3 scores for ERA and Systemic were significantly different from other subtypes, $P = 4.61 \times 10^{-3}$ and $P = 1.96 \times 10^{-8}$ respectively, matching the significant difference observed when performing a similar analysis using summary PC scores (Section 4.9.5).

4.10.4 Evaluating individual level eQTL data using the basis

In the previous section I outlined a method for computing posterior odds ratios given the genotypes for a set of individuals. This approach extends the basis considerably, allowing the projection of any measured phenotype for which raw genotype data is available. An obvious target are eQTL studies in relevant tissue types where genotypes and gene expression data are available. I wished to explore whether projecting the genotypes for such a study might have utility in assigning genes to basis PCs, and therefore provide insights into IMD disease pathogenesis, not achievable through the previously proposed approaches.

I obtained raw genotype and expression data from Raj et al. (2014), a published eQTL study of $CD4^+$ T cells, and $CD14^+$ monocytes; immune cell subtypes that have been implicated in IMD pathogenesis. In order to match the ancestry of projected genotypes with cohorts used to construct the basis, I took forward data only for individuals with American-European ancestry from Raj et al. (2014) for use in downstream analysis. After genotype harmonisation with the basis I computed posterior log odds ratios across all basis SNPs for 209 and 211 individuals with Monocytes $CD4^+$ T cells expression data respectively. In total I included 178 individuals for which expression values were available for both cell types. I began by projecting posterior odds ratios for each individual onto the basis obtaining a total of 2,343 PC scores across the eleven basis PCs. In order to assess whether there was a relationship between an individuals score and gene expression I fitted a linear regression model for each of the 19,323 genes whose expression was measured in the study, such that

$$Y \sim \beta_0 + \beta_1 X_k + \varepsilon, \tag{4.31}$$

where Y is a vector of normalised expression values for a given gene and cell type across all individuals and X_k is a corresponding vector of PC scores for the k^{th} component, β_0, β_1 and ε are intercept, coefficient and normal error terms ($\varepsilon \sim N(0, \sigma_\varepsilon^2)$) respectively. Such an approach resulted in the computation of 212,553 linear models across all 11 basis PCs. For each linear model the significance of the β_1 coefficient can be assessed using a t-statistic, under the null hypothesis that $\beta_1 = 0$ (i.e. there is no relationship between individual component score and gene expression). Using this approach I found that across PCs for both cell types there was little support for the association of gene expression for either cell type across any basis PC component. Indeed, at FDR<5% only two genes were

significant across all cell types and PCs examined (Table 4.6), neither of which shows evidence for involvement in IMD aetiology.

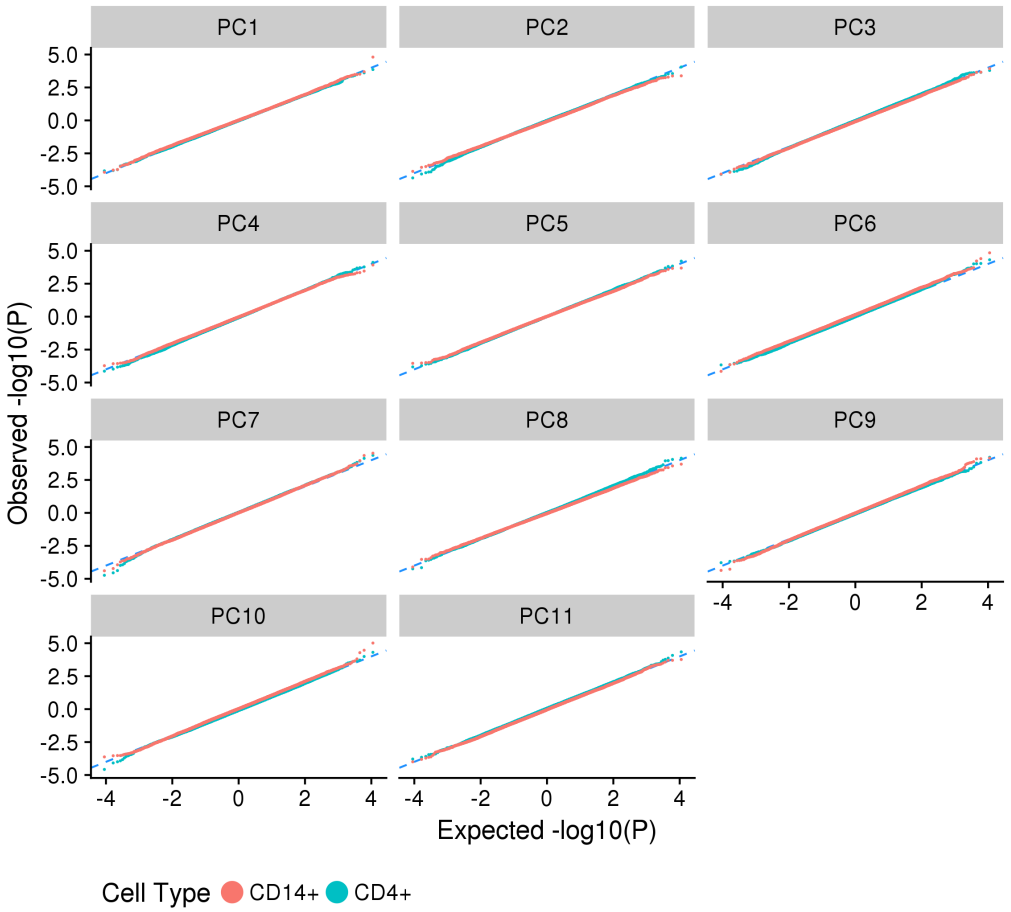


Fig. 4.25 Quantile-quantile normal plots for *p*-values obtained from regressing an individual's PC score with expression data from Raj et al. (2014). Colours represent cell-types; $CD14^{+}$ (Monocytes) - coral and $CD4^{+}$ - turquoise.

Whilst such an observation is disappointing, there could be a number of explanations for this. Firstly, variants that effect gene expression are nominally found in proximity (*cis*) to their target gene (Brem et al., 2002; GTEx Consortium, 2017) and whilst evidence for *trans*-eQTLs are emerging, their robust detection requires sample sizes 200 times larger than that analysed here (Võsa et al., 2018). This means that for a given gene a majority of input SNPs will be non-informative, instead adding stochastic noise, and it is likely that this noise will be sufficient to overcome any true signal arising from variants with genuine associations with expression.

PC	Probe	Gene	$\hat{\beta}_1$	p	p_{adj}
PC6	8073207	*FAM83F*	30.22	2.5×10^{-6}	0.048
PC10	8152280	*LRP12*	49.25	1.1×10^{-6}	0.022

Table 4.6 Gene expression from Võsa et al. (2018) significantly associated with a basis principal component. PC - The basis principal component associated. $\hat{\beta}_1$ - regression coefficient estimate, p - raw p-value, p_{adj} - Benjamini-Hochberg adjusted p-value

4.11 Discussion

In this chapter I have proposed a PCA framework to construct a basis that can capture, in a summarised form, the shared and distinct genetic architectures across a set of input diseases using summary GWAS statistics. The comparison of GWAS summary statistics between related diseases is not novel (Cho and Feldman, 2015; Cotsapas et al., 2011; Gutierrez-Arcelus et al., 2016), but is commonly conducted by examining individual associations signals between between pairs of disease associations. Such approaches are attractive as they simplify the comparison of effect sizes and putative causal variants which in turn can be used to suggest causal mechanisms. Genome-wide approaches such as genetic correlation can identify relationships between diseases (Bulik-Sullivan et al., 2015) but they do so in a pairwise fashion, and can miss relationships in the presence of opposing balanced correlation structures (Section 4.2). In contrast, the unsupervised approach presented here partitions and summarises the relationships between diseases in an holistic fashion, analysing the relationships between all available basis disease associations simultaneously. Whilst this learned basis is of interest, the projection of other traits into basis space is of great utility as it allows not only the characterisation of specific PCs that in turn can suggest biological insights, but also the partitioned (by PC) comparison between traits.

The application of PCA to feature extraction and dimension reduction over GWAS summary statistics has been attempted previously (Chang and Keinan, 2014), however my approach is markedly different in that I considered effect size and direction. I found that re-weighting input odds ratios prior to PCA was of great importance and failure to do so created a basis that was overwhelmed by stochastic noise at unassociated variants or possibly through more systematic noise due to study specific signals unrelated to underlying disease genetic architectures. The PCA of summary statistics for 10 IMDs that used a novel re-weighting method that I developed was able to discriminate between auto-inflammatory and autoimmune

diseases, in agreement with a clinical taxonomy suggested by (McGonagle and McDermott, 2006), providing support for the efficacy of the approach.

The basis obtained forms a compact summary of the relationships between input diseases The projection of external data onto the basis is especially useful as it allows the transfer of knowledge derived from diseases, with large sample size and resultant power, onto other phenotypes, which themselves can be, of very limited sample size. I found this particularly useful for biologically annotating individual PCs using UKBB data over a wide range of traits. Reassuringly, PC scores between UKBB SRD and matched basis diseases were most similar. Furthermore UKBB SR medication data showed a strong relationship with the diseases and corresponding treatments. I note that these results are concordant with the observations of Wu et al. (2019), who for example, found a strong relationship between RA and immuno-suppressant drugs using a polygenic risk score (PRS) approach.

On extending the approach to consider the more quantitative, intermediate phenotypes of blood count and eQTL data from Astle et al. (2016) and Võsa et al. (2018) respectively, my results were less clear. In comparison to the Mendelian randomisation analysis results described by Astle et al. (2016), I found only modest overlap.

For my analysis involving blood eQTLs (Võsa et al., 2018) I was limited somewhat by data availability, as summary statistics were only available across all genes for a limited subset of SNPs. To overcome this I made the assumption that effect sizes at missing SNPs were zero, and whilst this is a strong assumption, its effect can be somewhat ameliorated by considering a within study empirical significance measure. Whilst I was able to find significant PC scores for genes and pathways that corresponded to basis and other projected traits (e.g. integrin cell surface interactions and IBD (de Lange et al., 2017) and PC1) many were associated with more than one PC, mirroring the situation previously described where biological concepts are shared across multiple PCs. One possibility is that PCs are capturing different facets of the same biological concept (e.g. B-cell receptor activation pathway) with opposing effects across PCs and therefore diseases, however further work is required to describe this fully. Further work might consider alternative methods to achieve sparser PCs, such that biological concepts are concentrated across fewer components.

I examined in detail the application of the basis to provide insight into the genetic architecture of basis-related diseases, that due to their rarity or clinical heterogeneity, might lack sufficient power. Using JIA as example to investigate

this, I projected disease subtypes onto the basis, hypothesising that despite the small input sample size a PC score comparison might shed light as to whether genetic architecture might be responsible for the observed clinical heterogeneity. PC3 δ values were able to effectively discriminate Sys and ERA, which were statistically indistinguishable from the basis 'control' from other JIA subtypes, which all had significant negative δ values. This was unlikely to be completely driven by sample size due to the significance of PsA, the smallest subtype cohort. Whilst the studies of Hinks et al. (2017) and Ombrello et al. (2017) provide some support for the observed genetic heterogenity between subtypes.

In light of the projection of UKBB traits, which suggest that PC3 is a more general component for which a majority of diseases except Asthma and MS have significant δs. This is of some interest as PC3 scores for traits suggested a link between the efficacy of anti-TNF therapy. Whilst the effect of rs1800693 is well documented as a major factor in the exacerbating disease symptoms in MS (Gregory et al., 2012), via splicing of *TNFRSF1A*, this was not responsible for this observation as basis loadings for SNPs were minimal for the region 12p13.31 (median PC3 loading rank across all SNPs in the region was 210,673). Such an example provides an illustration as to how the approach could provide insight into disease pathology, although further work would be required to understand how robust such assertions might be.

I extended the framework to allow the projection of individual level genotype data into the basis space, using a Bayesian framework. This projection of posterior $\hat{\beta}$ onto the basis has overlap with the calculation of single trait polygenic risk scores (PRS) (Chatterjee et al., 2016), where rather than weighting allele counts by raw (or regularised) effect sizes ascertained from a single trait 'discovery' GWAS, posterior log odds ratios are weighted by the loadings obtained from the basis. In this way we obtain multi-trait PRS stratified by PC for each individual projected. The results from the application of this method to individual genotype data from the JIA disease subtype cohort, matched those generated using summary GWAS data, providing support for the framework. However on application to an eQTL dataset of relevant tissues, I was unable to find any convincing relationships between individual-level PC score and gene expression, even in light of the multiple examples of colocalisation between eQTL datasets and IMD GWAS signals that have previously been described in overlapping tissue contexts (Guo et al., 2015; Zhang et al., 2015). Although this appears contradictory to recent studies that highlight a significant role for *trans*-eQTLs (Boyle et al., 2017; Liu et al., 2018; Võsa et al., 2018), it seems most likely that the main reason I was

unable to confirm this was because of lack of power in the Raj et al. (2014) data set used. Further work will be required to examine whether larger studies eQTL studies in relevant tissues are able to provide more insight into PC characterisation.

One important consideration was the best approach to missing data, that arises from the heterogeneity of genotyping platforms and downstream imputation across SNPs with the GWAS I wished to consider. Whilst there are methods to deal with missing data using PCA, these are often computationally taxing and assume that data is missing at random (Ilin and Raiko, 2010). I chose to include a wider range of basis traits at the cost of decreased SNP coverage. I made this choice because I felt this would lead to a richer basis under the expectation that phenotypic heterogeneity is driven by differences in underlying genetic architecture. There is evidence to support this, for example T1D, where SNPs proximal to genes involved in pancreatic β-cell function (e.g. *GLIS3*, *BCAR1* and *INS*) are shared with type 2 diabetes but not other IMDs (Aylward et al., 2018; Fortune et al., 2015).

Whilst there are a myriad of further extensions that might be applied to the approach described, one of the most obvious is extending the number of traits used either to create or be projected on to the basis. For the former, given that summary statistics for well powered GWAS for a range of phenotypes are now routinely available (MacArthur et al., 2017), the possibility of creating a richer basis encompassing a much wider range of diseases and phenotypes that are not limited to IMD is possible. Such a resource might have utility in not only partitioning diseases, through PCs related to specific genetic architectures but also framing diseases in their overall context. This latter application might have utility in the development of new disease taxonomies that are based on molecular rather than clinical phenotypes. In addition to creating a richer basis, the identification and projection of additional rare forms of IMD holds promise. This could be extended over rare disease subtypes for which data is available, for example teasing apart antibody specific forms of eosinophilic granulomatosis with polyangiitis (Lyons et al., 2018) and perhaps using the basis to better understand the genetic architecture of even rarer diseases of the immune system where roles for both common and rare variants are suggested, such as Primary Immunodeficiency (Thaventhiran et al., 2018).

Chapter 5

Discussion

In this thesis I have developed methods for integrating genomic datasets at different resolutions in order to develop a better understanding of immune-mediated disease susceptibility. The COGS method (Chapter 3), concerned with suggesting causal variants, genes and tissue contexts using PCHi-C data and GWAS summary statistics, operates at the highest resolution, making use of the single base-pair resolution of SNPs. In contrast *blockshifter* (Chapter 2) acts at a lower resolution, employing a genome-wide approach, in order to suggest relationships between the disease specific genetic architectures and overall tissue specific, three dimensional chromatin organisation. Chapter 4 takes this further, describing a framework acting at a genome-wide and multi-trait scale, initially to summarise the genetic relationships between a set of clinically related diseases, but which has relevance to understanding the genetic basis of a wide range of both binary and quantitative human phenotypes.

5.1 Linking themes

5.1.1 Effect of single causal variant assumptions

Common to all three chapters is the use of Bayesian fine-mapping methods to assign posterior probabilities for a variant to be causal, predicated on the assumption of a single causal variant within an associated locus. In practice it is likely that this key assumption is often violated, and results from Asimit et al. (2019) suggest a sobering picture where the presence of multiple disease causing variants within a single locus might be currently underestimated.

The effect of violating this assumption is dependent on the locus and disease specific genetic architectures, but I envisage two main scenarios. In the most benign scenario, consider a locus in which there are multiple causal variants in low LD, but with heterogeneous effect sizes such that one 'lead' variant possesses a much larger association signal than the others. Setting aside technical reasons, such a situation could arise if the variant has a higher effect size or is more common in the population than the other causal variants. In such a case it is likely that the majority of the causal variant posterior probability is assigned to this 'lead' variant, thus masking the other causal variants in the locus.

As described in Asimit et al. (2019), a more worrying but rarer scenario consists of two causal variants, that whilst not being highly correlated with each other, are in LD with a third variant that itself is not causal. As, via LD, this third variant captures both causal variant associations, under single causal variant assumptions, it will misleadingly be assigned a high posterior probability for causality.

Such scenarios are of most relevance to the methods developed in Chapter 3 that operate at the highest resolution. The effect of the first scenario on COGS prioritisation will likely be more benign, as the prioritisation will be performed on the 'lead' signal at the expense of other causal variants within the locus. As a result genes and tissues, whilst of relevance to disease biology, are possibly incomplete, obscuring more subtle causal mechanisms. In contrast, scenario two is likely to result in a misleading prioritisation, as posterior probability is assigned to variants which are unrelated to disease pathogenesis, and as such could lead to genes and tissues unsuitable for informing downstream functional studies. Although this effect is likely to be mitigated by the size of the underlying *Hin*dIII fragments and the correlation between the states of neighbouring fragments. Whilst the ability to resolve these scenarios, once the single causal variant assumption is relaxed, is linked to underlying sample size, it might provide an explanation for the modest overlap of prioritised genes obtained for single and multiple causal variant fine-mapping inputs to COGS that I observed.

I expect the single causal variant assumption to have a more limited effect on the other, lower resolution approaches presented. Considering *blockshifter*, I would argue that this limitation although important at individual loci is somewhat averaged out at the genome-wide scale, and this is further ameliorated by the competitive nature of the test, where enrichment is computed not with respect to a null distribution, but as the comparison of two sets of tissues, similarly affected by underlying assumptions.

Whilst the shrinkage method underpinning the analyses in Chapter 4 is synthesised from the single causal variant fine-mapping approach, its goals are somewhat different in that it seeks to amplify effect sizes for SNPs which are likely to associated with disease whilst attenuating those that are more likely due to more stochastic processes. In such a situation picking a non-causal variant that encapsulates information about a multiple number of causal variants does not significantly disadvantage the method, under the assumption that input studies of shared genetic ancestry are selected. Indeed, for this application, where SNP coverage is limited it is more likely that a second key assumption, that the causal variant(s) are genotyped, is most likely violated, which I discuss in the next section.

5.1.2 Data availability

Another theme that runs through this thesis is that of data availability, mainly pertaining to GWAS datasets. This can be subdivided into two separate considerations: the public availability of full GWAS summary statistics for published studies, and due to the rapid technological advances in this area, the heterogeneous nature of SNP coverage between different traits.

With regards to the public availability of GWAS summary statistics, great strides have been made over the time since the commencement of this work. For example the work detailed in Chapters 2 and 3, required a considerable effort in order to compile a set of summary statistics over 31 traits. This fractured availability of GWAS summary statistics is in part historical and founded on somewhat overblown concerns about patient anonymity (Homer et al., 2008). Thankfully we are reaching a stage where online resources that aggregate and curate and make publicly available, full GWAS summary statistics across the whole range of human quantitative traits and disease phenotypes are being realised. For example The GWAS Catalog[1] (MacArthur et al., 2017), has taken this community requirement seriously and now provides access to the full summary statistics of over 750 studies/traits. However, given that in total they have curated data information about 'lead' SNPs for nearly 4,000 publications, clearly further work is required.

Critical to such efforts are not only suitable funding mechanisms that recognise the community benefit of such resources, but also the support of funders and peer-reviewed publications to encourage researchers to submit full summary statistics repository as part of the manuscript submission process. Such an approach,

[1]https://www.ebi.ac.uk/gwas/home

exemplified by ArrayExpress (Kolesnikov et al., 2015) and GEO (Clough and Barrett, 2016) for gene expression data, would not only provide better coverage of a diverse range of phenotypes but would present additional opportunities, for example they might provide a starting point for the development of light-touch meta-data standards for efficiently describing GWAS studies. Such standards need not be onerous but, for example, could incorporate guidelines on reporting effect sizes and the allele to which they pertain, enabling researchers to rapidly examine heterogeneity of effect direction at a given SNP or locus.

The differential SNP coverage between studies, presents a more technical challenge, and for the chapters concerned with PCHi-C integration a quote attributed to Charles Babbage springs to mind:

> Errors using inadequate data are much less than using no data at all.

Here 'inadequate data' is relevant to the use of the PMI methodology I developed in order to fill in instances where variants were not genotyped in the original study, and as the sobriquet suggests, this imputation method should not be conflated with the more rigorous procedures used to infer missing genotypes (Howie et al., 2009) or association statistics (Pasaniuc et al., 2014). Such an approach was necessary, as not only was the fine-mapping procedure I employed predicated on coverage of the causal variant, but also such resolution is required when integrating genomic annotations for reasons of sufficient coverage of individual features. For *blockshifter* I was able to simulate the effect of PMI on outcome, demonstrating that it decreased power to detect true enrichment, such that its effect on downstream results would be relatively benign. The affect on COGS prioritisation was more complicated and I showed that PMI, for certain loci could have marked effects on gene prioritisation, and in my final analysis I opted to use only dense ImmunoChip summary statistics, or those studies of IMD for which conventional imputation was performed (Bentham et al., 2015; Okada et al., 2014).

For the PCA framework, I argued that, given the design of genotyping platforms, the effect of a sparse SNP map on my results would be minimal as most common variation would be effectively tagged. This tension between data availability and the desire to cover as rich a set of diseases/phenotypes as possible is relevant throughout this thesis. Whilst a researcher, invested in a particular disease or phenotype might find this unacceptable, I took the view that covering as many traits as possible would lead to greater insight. On reflection this is somewhat application specific. It is reasonable to cast the net widely when attempting to elucidate relevant tissue contexts, or genetic relationships between diseases, whereas

a greater focus on specific diseases and fine-mapping strategies is reasonable when attempting a more detailed prioritisation of potential causal mechanisms using PCHi-C facilitated COGS at a specific locus.

5.1.3 The importance of orthogonal functional evidence

Undoubtedly, whilst the use of PCHi-C data in this thesis is novel there are multiple technical and biological reasons why it is no panacea for the elucidation of causal mechanisms in complex disease. A chief limitation is that resolution is restricted to the restriction fragment level and thus far removed from the theoretical single base-pair scale of genetic association studies. This limitation not only precludes the identification of causal variants but also introduces the possibility that multiple gene regulatory elements, acting on different target genes and in heterogeneous tissue contexts are contained within a single fragment. Leaving aside the technical challenges of employing alternative, 'frequent-cutter' restriction enzymes strategies, an alternative method to overcome such challenges is to integrate orthogonal high resolution genomic data in an attempt to triangulate promising regulatory sequences containing putative causal variants. In this thesis I used an informal hierarchical method, first using COGS to prioritise genes, before descending to the individual promoter interacting fragment level to look for overlap with tissue matched annotations such as eRNAs, using this as a prospective filter to suggest promising candidates for functional followup (Table B.2). Clearly given the investment required in such downstream studies the examination of all sources of available evidence is prudent. Indeed in a number of regions LD prevents the statistical identification of causal variants, whilst large sample sizes and/or GWAS studies across different ethnicities has the potential to overcome this situation, such approaches are resource intensive. Thankfully, new high-throughput functional screens requiring only few donors are being developed to overcome some of these situations. For example massively parallel reporter assays (MPRA) (Melnikov et al., 2012) have been used successfully to interrogate the enhancer potential of regulatory sequences containing putative causal SNPs for red blood cell traits (Ulirsch et al., 2016). However such approaches are technically challenging and their reliance on plasmid transfection currently precludes the assessment of certain IMD tissue contexts such as monocyte lineages. Another technique which shows great promise is the use of CRISPR activation assays; where a strong transcriptional activator such as VP64 is tethered to a catalytically dead *Cas*9 which in conjunction with specific guide-RNAs can be targeted to

specific regulatory sequences. The motivation for such manipulations is they allow the assessment of the activity of a regulatory sequence in a tissue-context agnostic manner, as the proximal localisation of the transcriptional activator to the regulatory element will drive expression of the element and its target gene, irrespective of cellular context. Such an approach has been used to interrogate the *IL2RA* region, discussed in detail within this thesis, corroborating the contribution of rs61839660 to IMD disease risk discussed in Chapter 3 (Simeonov et al., 2017).

5.1.4 A new taxonomy

The importance of taxonomies for organising data in the biological sciences have been recognised since, 1758, when Carl Linnaeus proposed a systematic classification of plants and animals based on the detailed observations of many different organisms. Indeed, disease taxonomies such as The International Classification of Diseases (ICD) are the very heart of modern medicine, however these are beginning to lag behind a growing body of molecular phenotypes, obscured by similarities in clinical disease presentation, that can be important for efficiently treating disease (Mirnezami et al., 2012). To address this, new initiatives are being developed that can capture such rich, molecular phenotypes, in an extensible and flexible manner. One example is the Human Phenotype Ontology (HPO) (Köhler et al., 2017), which, whilst initially targeting rare disease, is being extended to cover more common diseases, providing not only detailed clinician-lead taxonomies but also facilitating relationships across disease domains. The development of such new taxonomies is intimately linked to the elucidation of the molecular phenotypes that underlie disease susceptibility as these will ultimately determine the spectrum of disease covered by a particular classification. The methods and analysis presented in Chapters 2 and 3 can be viewed as very early stage prioritisation methods for uncovering such molecular phenotypes by suggesting promising genes, tissue contexts and biological pathways that through careful functional validation, might lead to novel disease classifications.

Chapter 4 presents a promising approach for disease classification based on shared and distinct genetic architectures. In application to IMD, what emerges is a spectrum of risk encompassing auto-inflammatory diseases such as IBD at one extreme and classic autoimmune diseases such as SLE at the other. The presence of blurred lines between different but clinically related diseases is not unexpected (Cleynen et al., 2016; McGonagle and McDermott, 2006), however the ability to 'learn' orthogonal partitions of risk across common clinically related

diseases affords the opportunity to project rarer phenotypes in order to uncover unappreciated genetic similarities and differences. Whilst this is of some interest, characterising the biology that each component relates to will greatly increase utility of the method, whether for suggesting novel therapeutic avenues (or those to avoid) or to realise a deeper molecular classification of particular disease. However the approach, as presented here, suffers somewhat from the overlap of concepts between components which makes relating them to biological processes challenging. Some of this overlap could be biological, for example the same variants in a particular biological pathway exerting antagonistic effects in different diseases and further work looking at PC loadings and variants will be required to investigate this.

A natural extension is to consider methods, such as VARIMAX (Jackson, 2014), for transforming the principal components in order to promote sparsity in underlying loadings, which might result in greater discrimination of concepts across components. In practice given that the total number of components is small ($n = 11$) it is unlikely that this will have a profound effect, given that orthogonality, and thus variance explained, between components is to be preserved. It might be promising to investigate alternative matrix factorisation methods, where underlying assumptions of orthogonality are relaxed. Given the computational burdens of such techniques I would expect a filtering of SNPs, based on weightings would be propitious prior to their employment.

5.2 Further Work

5.2.1 PCHi-C facillitated gene prioritisation in alternative contexts

In chapter 3 I prioritised putative causal genes and tissue contexts based on PCHi-C maps of 17 haematopoietic cell types. As detailed above, there were a number of limitations both at the level of causal variant identification and PCHi-C resolution, and I outlined the importance of using multiple, orthogonal lines of evidence to assess prioritised causal mechanisms, prior to functional validation. One extension that might improve the robustness of such a prioritisation would be the incorporation of such alternate sources of evidence directly into the COGS methodology. Given its Bayesian foundations a promising avenue for future extensions could be the consideration of more informative priors, based on the

intersection of orthologous genomic datasets with GWAS association signals. As previously introduced in Chapter 2, *fgwas* (Pickrell, 2014) uses such an approach in order to suggest novel associations as well as identify genomic annotations of interest, and involves the computation of such informative priors simultaneously across multiple annotations for a given trait or disease. This invites a combined approach where *fgwas* is used to compute fine-mapping posterior probabilities in the context of relevant genomic annotations (e.g. ChIP-Seq), which are then used as an input to COGS to suggest causal genes. Such an analysis could provide pilot data indicating whether additional effort to mediate, with functional data, the beta binomial priors adopted in the multi-causal variant GUESSFM (Bottolo and Richardson, 2010; Wallace et al., 2015) approach would be worthwhile.

Another promising application of PCHi-C is in the identification of causal mechanisms underlying monogenic causes of disease. As detailed in the Chapter 1 rare variation in non-coding regions can result in profound phenotypic effects (Lettice et al., 2003). Given the availability of whole genome sequences (WGS) and extensive phenotypic information for a wide range of rare diseases (Ouwehand, 2019), PCHi-C could be applied in a number of ways in order to further understand the causal mechanisms underlying rare diseases. For example WGS data could be mined for the presence of compound heterozygotes, where PCHi-C data is used to 'connect' a heterozygous regulatory sequence deletion with a deleterious protein truncating single nucleotide variant in the same gene. Alternatively, if a rare disease is characterised by specific tissue defects (especially those with relevance to Haematopoietic cell lineages), tissue-specific COGS scores from related complex diseases (e.g. IMD and primary immune deficiency) could be used to suggest causal genes underlying the rare disease. Here the notion is that both rare and common diseases share causal genes, but differ in the level of penetrance and expressivity. Such applications, whilst in their infancy are beginning to bear fruit (Thaventhiran et al., 2018).

5.2.2 Further exploration of basis polygenic risk scores

With the advent of genomic technologies, we are now able to measure, with unprecedented resolution and accuracy, molecular phenotypes in a scalable fashion, ushering in an era of 'genomic medicine'. Genetics is forming the vanguard of such an approach, through the development of polygenic risk-scores (PRS), which, in their simplest manifestation, are formed from a weighted sum of a set of risk variants present within an individual (Chatterjee et al., 2016; Wray et al., 2007).

Such approaches are not new and have been applied in the field of livestock breeding for many years, but, with the advent GWASs of large sample size, are now being applied to human disease (Wray et al., 2019). In certain settings, for example breast cancer, PRS, are beginning to be assessed for their utility in augmenting more traditional risk profiling measures such as family history and highly penetrant mutations in *BRCA1/2* so as to provide a clinical screening mechanism for identifying individuals at high risk of developing disease and whether this might manifest as a particularly aggressive disease subtype (e.g. oestrogen receptor negative breast cancers) (Mavaddat et al., 2019). Setting aside technical issues, for example, the applicability of such GRS across different populations, their translation into a clinical setting is challenging. Firstly, the predictive power of a PRS is tied to the narrow-sense heritability of the trait, which for example in breast cancer is approximately 40%, but can extend to 90% for example, in ankylosing spondylitis. Even in such diseases which are highly heritable, their application could lead to a prevention paradox (Rose, 1981), where only a small number of cases come from high risk individuals where such a PRS might be discriminatory, thus undermining their use in a clinical setting. Ultimately, the clinical utility of a disease-specific PRS, will be shaped by the existence and availability of interventions that modulate risk in individuals identified as being at high risk. For example knowing that a young child is at high genetic predisposition of being obese is likely to have more clinical utility than such an observation in an adult patient (Hunter and Drazen, 2019).

In my opinion for the reasons set out above, the applicability of PRS to individual disease prediction/screening in a clinical setting for IMD is somewhat limited, and their maximum utility most probably lies elsewhere, for example in disease stratification. In such a setting individuals are stratified as to their risk of developing a particular IMD, and high risk individuals are then targeted for detailed longitudinal study. In chapter 4, I made a brief reference to the parallels between such PRS and the PCA basis developed. In the context of individual genotype projections, basis PC scores are derived from linear combinations of individual allele counts giving an individual PC score for each component and can thus be interpreted as partitioned multi-trait PRS. Further exploration of the presented method is, I think justified, given the results observed for JIA subtypes. One potential application might be to use the basis to interrogate disease prognosis in Crohn's disease, as a previous within-cases GWAS presented evidence such a process might be orthogonal to disease susceptibility (Lee et al., 2017), and further analysis might yield additional, clinically relevant insights. The benefit of

this cohort is that participants were recruited at phenotypic extremes under the expectation this would maximise genetic discrimination, and as such might exist at the extremes of the distributions for polygenic risk. The projection of both summary and individual data from this study on to the basis thus framing them in the context of multiple IMDs, could help to shore up such evidence of orthogonality, and also lead to additional insights into previously unexamined overlaps in IMD genetic architectures with disease prognosis. One further application might be to use the basis to further explore the genetic architecture of rare diseases, that manifest an immune-mediated component, for example common variable immune deficiency (Li et al., 2015). Such an analysis could suggest, not only common molecular aetiologies with prexisting IMDs, but might give some insight into the interplay between common and rare variants underlying disease pathogenesis.

5.3 Concluding Remarks

For a majority of common human disease we are now firmly in the 'post-GWAS' era. Whilst ever larger sample sizes, coupled with new sequencing technologies will extend the collection of associated loci and causal variants, focus is now turning towards elucidating underlying causal mechanisms (Visscher et al., 2017). As set out in this thesis and elsewhere identifying causal variants, their target effectors and relevant tissue contexts is challenging. Such an effort is justified, however as the rewards, are likely to be substantial, encompassing the identification of novel therapeutic targets, and the development of disease classifications built on firm genetic and molecular foundations. Only once these are realised will the large amounts of time and resources invested in GWAS over the last decade be fully justified.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. et al. (2015). A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Aitkin, M. and Chadwick, T. (2003). Bayesian analysis of 2 x 2 contingency tables from comparative trials.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., Hale, C., Dougan, G. and Gaffney, D.J. (2018, March). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics 50*(3), 424–431.

Albert, F.W. and Kruglyak, L. (2015, April). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics 16*(4), 197–212.

Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A. et al. (2011, March). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet. 43*(3), 246–252.

Aschard, H., Vilhjálmsson, B.J., Greliche, N., Morange, P.E., Trégouët, D.A. and Kraft, P. (2014, May). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *American Journal of Human Genetics 94*(5), 662–676.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000, May). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet 25*(1), 25–29.

Asimit, J.L., Rainbow, D.B., Fortune, M., Grinberg, N.F., Wicker, L.S. and Wallace, C. (2019, January). Sharing information between related diseases using Bayesian joint fine mapping increases accuracy and identifies novel associations in six immune mediated diseases. *bioRxiv*, 553560.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A. et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell 167*(5), 1415–1429.e19.

Avery, C.L., He, Q., North, K.E., Ambite, J.L., Boerwinkle, E., Fornage, M., Hindorff, L.A., Kooperberg, C., Meigs, J.B., Pankow, J.S. et al. (2011, October). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS genetics 7*(10), e1002322.

Aylward, A., Chiou, J., Okino, M.L., Kadakia, N. and Gaulton, K.J. (2018, November). Shared genetic risk contributes to type 1 and type 2 diabetes etiology. *Human Molecular Genetics*.

Bahcall, O. (2012, July). Asking for more. *Nature Genetics 44*(7), 733.

Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C. et al. (2009, June). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet. 41*(6), 703–707.

Bassil, R., Orent, W., Olah, M., Kurdi, A.T., Frangieh, M., Buttrick, T., Khoury, S.J. and Elyaman, W. (2014, July). BCL6 controls Th9 cell development by repressing Il9 transcription. *J Immunol 193*(1), 198–207.

Bell, G.I., Horita, S. and Karam, J.H. (1984, February). A Polymorphic Locus Near the Human Insulin Gene Is Associated with Insulin-dependent Diabetes Melliitus. *Diabetes 33*(2), 176–183.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and Speed, T.P. (2012, May). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research 40*(10), e72–e72.

Bentham, J., Morris, D.L., Graham, D.S.C., Pinder, C.L., Tombleson, P., Behrens, T.W., Martín, J., Fairfax, B.P., Knight, J.C., Chen, L. et al. (2015, December). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics 47*(12), 1457–1464.

Bickel, P.J., Boley, N., Brown, J.B., Huang, H. and Zhang, N.R. (2010, December). Subsampling methods for genomic inference. *The Annals of Applied Statistics 4*(4), 1660–1697.

Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.L.V. and Kruglyak, L. (2013, February). Finding the sources of missing heritability in a yeast cross. *Nature 494*(7436), 234–237.

Bonev, B., Cohen, N.M., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A. et al. (2017, October). Multiscale 3d Genome Rewiring during Mouse Neural Development. *Cell 171*(3), 557–572.e24.

Bottolo, L. and Richardson, S. (2010, September). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis 5*(3), 583–618.

Bousquet, J., Chanez, P., Lacoste, J.Y., Barnéon, G., Ghavanian, N., Enander, I., Venge, P., Ahlstedt, S., Simony-Lafontaine, J. and Godard, P. (1990, October). Eosinophilic inflammation in asthma. *The New England Journal of Medicine 323*(15), 1033–1039.

Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017, June). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell 169*(7), 1177–1186.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002, April). Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.) 296*(5568), 752–755.

Brewerton, D.A., Hart, F.D., Nicholls, A., Caffrey, M., James, D.C. and Sturrock, R.D. (1973, April). Ankylosing spondylitis and HL-A 27. *Lancet (London, England) 1*(7809), 904–907.

Buenrostro, J., Wu, B., Chang, H. and Greenleaf, W. (2015, January). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] 109*, 21.29.1–21.29.9.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L. et al. (2015, November). An atlas of genetic correlations across human diseases and traits. *Nature Genetics 47*(11), 1236–1241.

Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L. and Neale, B.M. (2015, March). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics 47*(3), 291–295.

Burren, O.S., Guo, H. and Wallace, C. (2014). VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics 30*(23), 3342–3348.

Burren, O.S., Rubio García, A., Javierre, B.M., Rainbow, D.B., Cairns, J., Cooper, N.J., Lambourne, J.J., Schofield, E., Castro Dopico, X., Ferreira, R.C. et al. (2017, September). Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biology 18*(1), 165.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.M., Osborne, C. et al. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol 17*(1), 127.

Castellanos-Rubio, A., Fernandez-Jimenez, N., Kratchmarov, R., Luo, X., Bhagat, G., Green, P.H.R., Schneider, R., Kiledjian, M., Bilbao, J.R. and Ghosh, S. (2016, April). A long noncoding RNA associated with susceptibility to celiac disease. *Science 352*(6281), 91–95.

Chang, D. and Keinan, A. (2014, September). Principal component analysis characterizes shared pathogenetics from genome-wide association studies. *PLoS computational biology 10*(9), e1003820.

Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity 56*(1-3), 18–31.

Chatterjee, N., Shi, J. and García-Closas, M. (2016, July). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature reviews. Genetics 17*(7), 392–406.

Chesi, A., Wagley, Y., Johnson, M.E., Manduchi, E., Su, C., Lu, S., Leonard, M.E., Hodge, K.M., Pippin, J.A., Hankenson, K.D. et al. (2018, August). Genome-scale Capture C promoter interaction analysis implicates novel effector genes at GWAS loci for bone mineral density. *bioRxiv*, 405142.

Cho, J.H. and Feldman, M. (2015, July). Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. *Nat. Med. 21*(7), 730–738.

Choung, R.S., Unalp-Arida, A., Ruhl, C.E., Brantner, T.L., Everhart, J.E. and Murray, J.A. (2017, January). Less Hidden Celiac Disease But Increased Gluten Avoidance Without a Diagnosis in the United States: Findings From the National Health and Nutrition Examination Surveys From 2009 to 2014. *Mayo Clinic Proceedings 92*(1), 30–38.

Choy, M.K., Javierre, B.M., Williams, S.G., Baross, S.L., Liu, Y., Wingett, S.W., Akbarov, A., Wallace, C., Freire-Pritchett, P., Rugg-Gunn, P.J. et al. (2018, June). Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nature Communications 9*(1), 2526.

Church, C., Lee, S., Bagg, E.A.L., McTaggart, J.S., Deacon, R., Gerken, T., Lee, A., Moir, L., Mecinović, J., Quwailid, M.M. et al. (2009, August). A Mouse Model for the Metabolic Effects of the Human Fat Mass and Obesity Associated FTO Gene. *PLOS Genetics 5*(8), e1000599.

Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V. et al. (2015, September). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med 373*(10), 895–907.

Clayton, D. and Leung, H.T. (2007). An R package for analysis of whole-genome association studies. *Human Heredity 64*(1), 45–51.

Cleynen, I., Boucher, G., Jostins, L., Schumm, L.P., Zeissig, S., Ahmad, T., Andersen, V., Andrews, J.M., Annese, V., Brand, S. et al. (2016, January). Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet (London, England) 387*(10014), 156–167.

Clough, E. and Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in Molecular Biology (Clifton, N.J.) 1418*, 93–110.

Coit, P., Jeffries, M., Altorok, N., Dozmorov, M.G., Koelsch, K.A., Wren, J.D., Merrill, J.T., McCune, W.J. and Sawalha, A.H. (2013, June). Genome-wide DNA methylation study suggests epigenetic accessibility and transcriptional poising of interferon-regulated genes in naïve CD4+ T cells from lupus patients. *Journal of Autoimmunity 43*, 78–84.

Colbert, R.A. (2010, August). Classification of juvenile spondyloarthritis: enthesitis-related arthritis and beyond. *Nature reviews. Rheumatology 6*(8), 477–485.

Cooper, G.S., Bynum, M.L.K. and Somers, E.C. (2009, December). Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *Journal of Autoimmunity 33*(3-4), 197–207.

Cooper, J.D., Simmonds, M.J., Walker, N.M., Burren, O., Brand, O.J., Guo, H., Wallace, C., Stevens, H., Coleman, G., Wellcome Trust Case Control Consortium et al. (2012, December). Seven newly identified loci for autoimmune thyroid disease. *Human Molecular Genetics 21*(23), 5202–5208.

Cooper, N.J., Wallace, C., Burren, O.S., Cutler, A., Walker, N. and Todd, J.A. (2017, April). Type 1 diabetes genome-wide association analysis with imputation identifies five new risk regions. *bioRxiv*, 120022.

Cordell, H.J., Han, Y., Mells, G.F., Li, Y., Hirschfield, G.M., Greene, C.S., Xie, G., Juran, B.D., Zhu, D., Qian, D.C. et al. (2015). International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun. 6*, 8019.

Cortes, A. and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Res Ther 13*(1), 101.

Cortes, A., Dendrou, C.A., Motyer, A., Jostins, L., Vukcevic, D., Dilthey, A., Donnelly, P., Leslie, S., Fugger, L. and McVean, G. (2017, September). Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nature Genetics 49*(9), 1311–1318.

Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J. et al. (2011, August). Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLoS Genetics 7*(8).

Crick, F.H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology 12*, 138–163.

Cronin, C.C., Feighery, A., Ferriss, J.B., Liddy, C., Shanahan, F. and Feighery, C. (1997, December). High prevalence of celiac disease among patients with insulin-dependent (type I) diabetes mellitus. *The American Journal of Gastroenterology 92*(12), 2210–2212.

Cullen, K.E., Kladde, M.P. and Seyfred, M.A. (1993, July). Interaction between transcription regulatory regions of prolactin chromatin. *Science 261*(5118), 203–206.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M. et al. (2016). Next-generation genotype imputation service and methods. *Nature Genetics 48*(10), 1284–1287.

Davies, J.O.J., Oudelaar, A.M., Higgs, D.R. and Hughes, J.R. (2017, February). How best to identify chromosomal interactions: a comparison of approaches. *Nature Methods 14*(2), 125–134.

Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R. and Hughes, J.R. (2016, January). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods 13*(1), 74–80.

Davison, L.J., Wallace, C., Cooper, J.D., Cope, N.F., Wilson, N.K., Smyth, D.J., Howson, J.M.M., Saleh, N., Al-Jeffery, A., Angus, K.L. et al. (2012). Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum. Mol. Genet. 21*(2), 322–333.

de Boer, J.D., Majoor, C.J., van 't Veer, C., Bel, E.H.D. and van der Poll, T. (2012, April). Asthma and coagulation. *Blood 119*(14), 3236–3244.

de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.G. et al. (2017, February). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics 49*(2), 256–261.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002, February). Capturing Chromosome Conformation. *Science 295*(5558), 1306–1311.

Demenais, F., Margaritte-Jeannin, P., Barnes, K.C., Cookson, W.O.C., Altmüller, J., Ang, W., Barr, R.G., Beaty, T.H., Becker, A.B., Beilby, J. et al. (2018, January). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics 50*(1), 42–53.

Dendrou, C.A., Cortes, A., Shipman, L., Evans, H.G., Attfield, K.E., Jostins, L., Barber, T., Kaur, G., Kuttikkatte, S.B., Leach, O.A. et al. (2016, November). Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Science Translational Medicine 8*(363), 363ra149.

Dendrou, C.A., Plagnol, V., Fung, E., Yang, J.H.M., Downes, K., Cooper, J.D., Nutland, S., Coleman, G., Himsworth, M., Hardy, M. et al. (2009, September). Cell-specific protein phenotypes for the autoimmune locus IL2ra using a genotype-selectable human bioresource. *Nat Genet 41*(9), 1011–1015.

Devlin, B. and Risch, N. (1995, September). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics 29*(2), 311–322.

DIAGRAM Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J. et al. (2014, March). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics 46*(3), 234–244.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012, April). Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature 485*(7398), 376–380.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C. et al. (2006, October). Chromosome Conformation Capture Carbon Copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research 16*(10), 1299–1309.

Dryden, N.H., Broome, L.R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I. et al. (2014, November). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Research 24*(11), 1854–1868.

Du, Q., Luu, P.L., Stirzaker, C. and Clark, S.J. (2015, April). Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics 7*(6), 1051–1073.

Dubois, P.C.A., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A.R., Adány, R., Aromaa, A. et al. (2010, April). Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet. 42*(4), 295–302.

Efron, B. (1979, January). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics 7*(1), 1–26.

Eicher, J.D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J.P., Leslie, R. and Johnson, A.D. (2015, January). GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Research 43*(Database issue), D799–804.

Eijsbouts, C.Q., Burren, O.S., Newcombe, P.J. and Wallace, C. (2019, January). Fine mapping chromatin contacts in capture Hi-C data. *BMC genomics 20*(1), 77.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T. et al. (2007, June). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature 447*(7146), 799–816.

Ernst, J. and Kellis, M. (2012, March). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods 9*(3), 215–216.

Ernst, J. and Kellis, M. (2015, April). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol. 33*(4), 364–376.

Estrada, K., Styrkarsdottir, U., Evangelou, E., Hsu, Y.H., Duncan, E.L., Ntzani, E.E., Oei, L., Albagha, O.M.E., Amin, N., Kemp, J.P. et al. (2012, May). Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet. 44*(5), 491–501.

Evangelou, M., Smyth, D.J., Fortune, M.D., Burren, O.S., Walker, N.M., Guo, H., Onengut-Gumuscu, S., Chen, W.M., Concannon, P., Rich, S.S. et al. (2014, December). A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genet. Epidemiol. 38*(8), 661–670.

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K. et al. (2012, December). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics 44*(12), 1336–1340.

Eyre, S., Orozco, G. and Worthington, J. (2017, July). The genetics revolution in rheumatology: large scale genomic arrays and genetic mapping. *Nature Reviews. Rheumatology 13*(7), 421–432.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S. et al. (2016, January). The Reactome pathway Knowledgebase. *Nucleic Acids Res 44*(D1), D481–D487.

Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O. and Knight, J.C. (2012, May). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet 44*(5), 502–510.

Farh, K.K.H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A. et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature 518*(7539), 337–343.

Fortune, M.D., Guo, H., Burren, O., Schofield, E., Walker, N.M., Ban, M., Sawcer, S.J., Bowes, J., Worthington, J., Barton, A. et al. (2015, July). Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics 47*(7), 839–846.

Fortune, M.D. and Wallace, C. (2018, May). simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics. *bioRxiv*, 313023.

Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R. et al. (2010, December). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet. 42*(12), 1118–1125.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Elliott, K.S., Lango, H., Rayner, N.W. et al. (2007, May). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science 316*(5826), 889–894.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. et al. (2009, November). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature 462*(7269), 58–64.

García Rodríguez, L.A., Ruigómez, A. and Panés, J. (2006, May). Acute gastroenteritis is followed by an increased risk of inflammatory bowel disease. *Gastroenterology 130*(6), 1588–1594.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014, May). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet 10*(5), e1004383.

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y. et al. (2011). New gene functions in megakaryopoiesis and platelet formation. *Nature 480*(7376), 201–208.

Gmuca, S., Xiao, R., Brandon, T.G., Pagnini, I., Wright, T.B., Beukelman, T., Morgan, E.M. and Weiss, P.F. (2017). Multicenter inception cohort of enthesitis-related arthritis: variation in disease characteristics and treatment approaches. *Arthritis Research & Therapy 19*(1), 84.

Gregory, A.P., Dendrou, C.A., Attfield, K.E., Haghikia, A., Xifara, D.K., Butter, F., Poschmann, G., Kaur, G., Lambert, L., Leach, O.A. et al. (2012, August). TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature 488*(7412), 508–511.

GTEx Consortium, G. (2017, October). Genetic effects on gene expression across human tissues. *Nature 550*(7675), 204–213.

Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J.A. and Wallace, C. (2015). Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum. Mol. Genet.*.

Gutierrez-Arcelus, M., Rich, S.S. and Raychaudhuri, S. (2016, March). Autoimmune diseases - connecting risk alleles with molecular traits of the immune system. *Nature Reviews. Genetics 17*(3), 160–174.

Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J. et al. (2007, May). Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genetics 39*(5), 638–644.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

Hayter, S.M. and Cook, M.C. (2012, August). Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmunity Reviews 11*(10), 754–765.

Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S.R., Bauerfeind, A. et al. (2010, September). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature 467*(7314), 460–464.

Higgs, D.R., Vernimmen, D., Hughes, J. and Gibbons, R. (2007). Using Genomics to Study How Chromatin Influences Gene Expression. *Annual Review of Genomics and Human Genetics 8*(1), 299–325.

Hinks, A., Bowes, J., Cobb, J., Ainsworth, H.C., Marion, M.C., Comeau, M.E., Sudman, M., Han, B., Immunochip, J.A.C.f., Becker, M.L. et al. (2017, April). Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases. *Annals of the Rheumatic Diseases 76*(4), 765–772.

Hinks, A., Cobb, J., Marion, M.C., Prahalad, S., Sudman, M., Bowes, J., Martin, P., Comeau, M.E., Sajuthi, S., Andrews, R. et al. (2013, June). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature Genetics 45*(6), 664–669.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A. and Noble, W.S. (2012, March). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods 9*(5), 473–476.

Holgate, S.T., Noonan, M., Chanez, P., Busse, W., Dupont, L., Pavord, I., Hakulinen, A., Paolozzi, L., Wajdula, J., Zang, C. et al. (2011, June). Efficacy and safety of etanercept in moderate-to-severe asthma: a randomised, controlled trial. *The European Respiratory Journal 37*(6), 1352–1359.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F. and Craig, D.W. (2008, August). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics 4*(8), e1000167.

Hormozdiari, F., Kichaev, G., Yang, W.Y., Pasaniuc, B. and Eskin, E. (2015). Identification of causal genes for complex traits. *Bioinformatics 31*(12), i206–i213.

Howie, B.N., Donnelly, P. and Marchini, J. (2009, June). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics 5*(6), e1000529.

Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C.A., Andersen, V., Cleynen, I., Cortes, A., Crins, F. et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature 547*(7662), 173–178.

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R. and Higgs, D.R. (2014, February). Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics 46*(2), 205–212.

Hunter, D.J. and Drazen, J.M. (2019, May). Has the Genome Granted Our Wish Yet? *The New England Journal of Medicine*.

ICBP, Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C. et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature 478*(7367), 103–109.

Ilin, A. and Raiko, T. (2010). Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research 11*(Jul), 1957–2000.

IMSGC, Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D. et al. (2013, November). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics 45*(11), 1353–1360.

IMSGC, Wellcome Trust Case Control Consortium 2, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C.A., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z. et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature 476*(7359), 214–219.

Inshaw, J.R.J., Cutler, A.J., Burren, O.S., Stefana, M.I. and Todd, J.A. (2018, July). Approaches and advances in the genetic causes of autoimmune disease and their implications. *Nature Immunology 19*(7), 674–684.

International Genetics of Ankylosing Spondylitis Consortium (IGAS), Cortes, A., Hadler, J., Pointon, J.P., Robinson, P.C., Karaderi, T., Leo, P., Cremin, K., Pryce, K., Harris, J. et al. (2013, July). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature Genetics 45*(7), 730–738.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P. et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature 449*(7164), 851–861.

Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., UK10K Consortium, Dunham, I., Birney, E. et al. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature Genetics 51*(2), 343–353.

Jackson, J.E. (2014). Varimax Rotation. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society.

Jacob, F. and Monod, J. (1961, June). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology 3*(3), 318–356.

Jameson, J.L. and Longo, D.L. (2015, June). Precision medicine–personalized, problematic, and promising. *The New England Journal of Medicine 372*(23), 2229–2234.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J. et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell 167*(5), 1369–1384.e19.

Jeffreys, H. (1973). *Scientific Inference*. Cambridge University Press.

Ji, S.G., Juran, B.D., Mucha, S., Folseraas, T., Jostins, L., Melum, E., Kumasaka, N., Atkinson, E.J., Schlicht, E.M., Liu, J.Z. et al. (2017, February). Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nature Genetics 49*(2), 269–273.

Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. et al. (2001, October). Haplotype tagging for the identification of common disease genes. *Nature Genetics 29*(2), 233–237.

Jones, P.A. (2012, July). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics 13*(7), 484–492.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A. et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature 491*(7422), 119–124.

Jäger, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H.E., Heindl, A., Whiffin, N., Carnicer, M.J., Broome, L., Dryden, N. et al. (2015, February). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications 6*, 6178.

Kanehisa, M. and Goto, S. (2000, January). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res 28*(1), 27–30.

Kanhere, A. and Bansal, M. (2005). Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research 33*(10), 3165–3175.

Kass, R.E. and Raftery, A.E. (1995, June). Bayes Factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kichaev, G., Yang, W.Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P. and Pasaniuc, B. (2014, October). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet 10*(10), e1004722.

Klei Lambertus, Luca Diana, Devlin B. and Roeder Kathryn (2007, October). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology 32*(1), 9–19.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T. et al. (2005, April). Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.) 308*(5720), 385–389.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T. et al. (2015, January). ArrayExpress update—simplifying data submissions. *Nucleic Acids Research 43*(D1), D1113–D1116.

Kumasaka, N., Knights, A.J. and Gaffney, D.J. (2019, January). High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nature Genetics 51*(1), 128–137.

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M. et al. (2017). The Human Phenotype Ontology in 2017. *Nucleic Acids Research 45*(D1), D865–D876.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P. et al. (2012, September). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research 22*(9), 1813–1831.

Lawrence, M., Daujat, S. and Schneider, R. (2016, January). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics 32*(1), 42–56.

Lee, J.C., Biasci, D., Roberts, R., Gearry, R.B., Mansfield, J.C., Ahmad, T., Prescott, N.J., Satsangi, J., Wilson, D.C., Jostins, L. et al. (2017, February). Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nature Genetics 49*(2), 262–268.

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R. et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics 50*(8), 1112–1121.

Lee, Y., Francesca, L., Pique-Regi, R. and Wen, X. (2018, May). Bayesian Multi-SNP Genetic Association Analysis: Control of FDR and Use of Summary Statistics. *bioRxiv*, 316471.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature 536*(7616), 285–291.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003, July). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics 12*(14), 1725–1735.

Li, J., Jørgensen, S.F., Maggadottir, S.M., Bakay, M., Warnatz, K., Glessner, J., Pandey, R., Salzer, U., Schmidt, R.E., Perez, E. et al. (2015). Association of CLEC16a with human common variable immunodeficiency disorder and role in murine B cells. *Nat. Commun. 6*, 6804.

Li, W., Notani, D. and Rosenfeld, M.G. (2016, April). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews. Genetics 17*(4), 207–223.

Li, Y.R., Zhao, S.D., Li, J., Bradfield, J.P., Mohebnasab, M., Steel, L., Kobie, J., Abrams, D.J., Mentch, F.D., Glessner, J.T. et al. (2015). Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nat. Commun. 6*, 8442.

Liao, W., Lin, J.X. and Leonard, W.J. (2013, January). Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity 38*(1), 13–25.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015, December). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst 1*(6), 417–425.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. et al. (2009, October). Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (New York, N.Y.) 326*(5950), 289–293.

Lin, D.Y. and Sullivan, P.F. (2009, December). Meta-analysis of genome-wide association studies with overlapping subjects. *American Journal of Human Genetics 85*(6), 862–872.

Lindor, K.D., Gershwin, M.E., Poupon, R., Kaplan, M., Bergasa, N.V., Heathcote, E.J. and American Association for Study of Liver Diseases (2009, July). Primary biliary cirrhosis. *Hepatology (Baltimore, Md.) 50*(1), 291–308.

Liu, J.Z., Mcrae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G. et al. (2010, July). A Versatile Gene-Based Test for Genome-wide Association Studies. *American Journal of Human Genetics 87*(1), 139–145.

Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T. et al. (2015, September). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics 47*(9), 979–986.

Liu, X., Li, Y.I. and Pritchard, J.K. (2018, September). Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv*, 425108.

Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J. et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature 518*(7538), 197–206.

Long, M.D., Martin, C.F., Pipkin, C.A., Herfarth, H.H., Sandler, R.S. and Kappelman, M.D. (2012, August). Risk of Melanoma and Nonmelanoma Skin Cancer Among Patients With Inflammatory Bowel Disease. *Gastroenterology 143*(2), 390–399.e1.

Lyons, P., Peters, J., Alberici, F., Liley, J., Coulson, R., Astle, W., Baldini, C., Bonatti, F., Cid, M., Elding, H. et al. (2018, December). Genetically distinct clinical subsets, and associations with asthma and eosinophil abundance, within Eosinophilic Granulomatosis with Polyangiitis. *bioRxiv*, 491837.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. et al. (2017, January). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research 45*(Database issue), D896–D901.

Manning, A.K., Hivert, M.F., Scott, R.A., Grimsby, J.L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.T., Bielak, L.F., Prokopenko, I. et al. (2012, June). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet. 44*(6), 659–669.

Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., Cooper, N.J., Barton, A., Wallace, C., Fraser, P. et al. (2015, November). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature Communications 6*.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. et al. (2012, September). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.) 337*(6099), 1190–1195.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.H., Wang, Q., Bolla, M.K. et al. (2019, January). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American Journal of Human Genetics 104*(1), 21–34.

McCarthy, M.I. (2017). Painting a new picture of personalised medicine for diabetes. *Diabetologia 60*(5), 793–799.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K. et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics 48*(10), 1279–1283.

McGonagle, D. and McDermott, M.F. (2006, August). A proposed classification of the immunological diseases. *PLoS medicine 3*(8), e297.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. and Cunningham, F. (2016, June). The Ensembl Variant Effect Predictor. *Genome Biology 17*, 122.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B. et al. (2012, February). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology 30*(3), 271–277.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A. et al. (2015, June). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics 47*(6), 598–606.

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Rolando, D.M.Y., Javierre, B.M., Atla, G., Farabella, I., Morgan, C.C. et al. (2018, August). Human pancreatic islet 3d chromatin architecture provides insights into the genetics of type 2 diabetes. *bioRxiv*, 400291.

Millard, L.A.C., Davies, N.M., Gaunt, T.R., Davey Smith, G. and Tilling, K. (2017, October). Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *International Journal of Epidemiology*.

Miller, A.J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 389–425.

Mirnezami, R., Nicholson, J. and Darzi, A. (2012, February). Preparing for Precision Medicine. *New England Journal of Medicine 366*(6), 489–491.

Morrell, C.N., Aggrey, A.A., Chapman, L.M. and Modjeski, K.L. (2014, May). Emerging roles for platelets as immune and inflammatory cells. *Blood 123*(18), 2759–2767.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A. et al. (2012, September). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet. 44*(9), 981–990.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J. and Chang, H.Y. (2016, November). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods 13*(11), 919–922.

Murphy, K. and Weaver, C. (2017). *Janeway's Immunobiology* (9th ed.). Garland Science. Taylor & Francis Group, LLC.

Márquez, A., Kerick, M., Zhernakova, A., Gutierrez-Achury, J., Chen, W.M., Onengut-Gumuscu, S., González-Álvaro, I., Rodriguez-Rodriguez, L., Rios-Fernández, R., González-Gay, M.A. et al. (2018). Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. *Genome Medicine 10*(1), 97.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S. et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature 506*(7488), 376–381.

Ombrello, M.J., Arthur, V.L., Remmers, E.F., Hinks, A., Tachmazidou, I., Grom, A.A., Foell, D., Martini, A., Gattorno, M., Özen, S. et al. (2017, May). Genetic architecture distinguishes systemic juvenile idiopathic arthritis from other forms of juvenile idiopathic arthritis: clinical and therapeutic implications. *Annals of the Rheumatic Diseases 76*(5), 906–913.

Onengut-Gumuscu, S., Chen, W.M., Burren, O., Cooper, N.J., Quinlan, A.R., Mychaleckyj, J.C., Farber, E., Bonnie, J.K., Szpak, M., Schofield, E. et al. (2015, April). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics 47*(4), 381–386.

Ouwehand, W.H. (2019, January). Whole-genome sequencing of rare disease patients in a national healthcare system | bioRxiv.

Parkes, M., Cortes, A., van Heel, D.A. and Brown, M.A. (2013, September). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews. Genetics 14*(9), 661–673.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D.P., Patterson, N. and Price, A.L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics 30*(20), 2906–2914.

Paul, D.S., Teschendorff, A.E., Dang, M.A.N., Lowe, R., Hawa, M.I., Ecker, S., Beyan, H., Cunningham, S., Fouts, A.R., Ramelius, A. et al. (2016, November). Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nature Communications 7*, 13555.

Pearson, K. and Blakeman, J. (1906). *On the Theory of Contingency and Its Relation to Association and Normal Correlation*. Number v. 1-4 in A Mathematical Theory of Random Migration. Dulau and Company.

Peters, J.E., Lyons, P.A., Lee, J.C., Richard, A.C., Fortune, M.D., Newcombe, P.J., Richardson, S. and Smith, K.G.C. (2016, March). Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS Genet 12*(3), e1005908.

Petty, R.E., Southwood, T.R., Manners, P., Baum, J., Glass, D.N., Goldenberg, J., He, X., Maldonado-Cocco, J., Orozco-Alcala, J., Prieur, A.M. et al. (2004, February). International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *The Journal of Rheumatology 31*(2), 390–392.

Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet. 94*(4), 559–573.

Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y. and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics 48*(7), 709–717.

Prahalad, S., Conneely, K.N., Jiang, Y., Sudman, M., Wallace, C.A., Brown, M.R., Ponder, L.A., Rohani-Pichavant, M., Zwick, M.E., Cutler, D.J. et al. (2013, June). Susceptibility to childhood-onset rheumatoid arthritis: investigation of a weighted genetic risk score that integrates cumulative effects of variants at five genetic loci. *Arthritis and Rheumatism 65*(6), 1663–1667.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006, August). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*(8), 904–909.

project consortium, U. (2015, October). The UK10k project identifies rare variants in health and disease. *Nature 526*(7571), 82–90.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. et al. (2007, September). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics 81*(3), 559–575.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics 47*, 11.12.1–11.1234.

Rainbow, D.B., Pekalski, M.L., Cutler, A., Burren, O.S., Walker, N., Todd, J., Wallace, C. and Wicker, L. (2017, January). A rare IL2ra haplotype identifies SNP rs61839660 as causal for autoimmunity. *bioRxiv*.

Rainbow, D.B., Yang, X., Burren, O., Pekalski, M.L., Smyth, D.J., Klarqvist, M.D.R., Penkett, C.J., Brugger, K., Martin, H., Todd, J.A. et al. (2015, November). Epigenetic analysis of regulatory T cells using multiplex bisulfite sequencing. *Eur J Immunol 45*(11), 3200–3203.

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I. et al. (2014, May). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science (New York, N.Y.) 344*(6183), 519–523.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. et al. (2014, December). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell 159*(7), 1665–1680.

Ravelli, A. and Martini, A. (2007, March). Juvenile idiopathic arthritis. *Lancet (London, England) 369*(9563), 767–778.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C.Y., Consortium, I.S., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D. et al. (2009, June). Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLOS Genetics 5*(6), e1000534.

Rice, G.I., Reijns, M.A., Coffin, S.R., Forte, G.M., Anderson, B.H., Szynkiewicz, M., Gornall, H., Gent, D., Leitch, A., Botella, M.P. et al. (2013, August). Synonymous mutations in RNASEH2a create cryptic splice sites impairing RNase H2 enzyme function in Aicardi-Goutières syndrome. *Human mutation 34*(8), 1066–1070.

Risch, N. and Merikangas, K. (1996, September). The future of genetic studies of complex human diseases. *Science (New York, N.Y.) 273*(5281), 1516–1517.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015, April). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research 43*(7), e47.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J. et al. (2015, February). Integrative analysis of 111 reference human epigenomes. *Nature 518*(7539), 317–330.

Rockman, M.V. (2012, January). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution; International Journal of Organic Evolution 66*(1), 1–17.

Rose, G. (1981, June). Strategy of prevention: lessons from cardiovascular disease. *British Medical Journal (Clinical Research Ed.) 282*(6279), 1847–1851.

Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Constortium, I.I.B.D.G., Cotsapas, C. and Daly, M.J. (2011, January). Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLOS Genetics 7*(1), e1001273.

Russo, R.A.G. and Katsicas, M.M. (2009, May). Clinical remission in patients with systemic juvenile idiopathic arthritis treated with anti-tumor necrosis factor agents. *The Journal of Rheumatology 36*(5), 1078–1082.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N. and Reich, D. (2014, March). The genomic landscape of Neanderthal ancestry in present-day humans., The landscape of Neandertal ancestry in present-day humans. *Nature 507*(7492), 354–357.

Schofield, E.C., Carver, T., Achuthan, P., Freire-Pritchett, P., Spivakov, M., Todd, J.A. and Burren, O.S. (2016). CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics (Oxford, England) 32*(16), 2511–2513.

Schork, A.J., Thompson, W.K., Pham, P., Torkamani, A., Roddey, J.C., Sullivan, P.F., Kelsoe, J.R., O'Donovan, M.C., Furberg, H., Consortium, T.T.a.G. et al. (2013, April). All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLOS Genetics 9*(4), e1003449.

Scuteri, A., Sanna, S., Chen, W.M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G. et al. (2007, July). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS genetics 3*(7), e115.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012, February). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell 148*(3), 458–472.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001, January). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research 29*(1), 308–311.

Shogren-Knaak, M., Ishii, H., Sun, J.M., Pazin, M.J., Davie, J.R. and Peterson, C.L. (2006, February). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (New York, N.Y.) 311*(5762), 844–847.

Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C. et al. (2018, March). *De novo* mutations in regulatory elements in neurodevelopmental disorders. *Nature 555*(7698), 611–616.

Simeonov, D.R., Gowen, B.G., Boontanrart, M., Roth, T.L., Gagnon, J.D., Mumbach, M.R., Satpathy, A.T., Lee, Y., Bray, N.L., Chan, A.Y. et al. (2017, September). Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature 549*(7670), 111–115.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006, November). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature Genetics 38*(11), 1348–1354.

Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F. et al. (2014, March). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature 507*(7492), 371–375.

Smyth, D.J., Plagnol, V., Walker, N.M., Cooper, J.D., Downes, K., Yang, J.H.M., Howson, J.M.M., Stevens, H., McManus, R., Wijmenga, C. et al. (2008, December). Shared and distinct genetic variants in type 1 diabetes and celiac disease. *The New England Journal of Medicine 359*(26), 2767–2777.

Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. (2013, July). Pleiotropy in complex traits: challenges and strategies. *Nature reviews. Genetics 14*(7), 483–495.

Somers, E.C., Thomas, S.L., Smeeth, L. and Hall, A.J. (2006, March). Autoimmune Diseases Co-occurring Within Individuals and Within Families: A Systematic Review. *Epidemiology 17*(2), 202.

Song, L. and Crawford, G.E. (2010, February). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols 2010*(2), pdb.prot5384.

Song, M., Yang, X., Ren, X., Wang, C., Jacob, F., Wu, K., Traglia, M., Li, B., Maliskova, L., Jones, I. et al. (2018, December). cis-Regulatory Chromatin Contacts in Neural Cells Reveal Contributions of Genetic Variants to Complex Neurological Disorders. *bioRxiv*, 494450.

Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M. et al. (2009, November). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics 41*(11), 1182–1190.

Soskic, B., Cano-Gamez, E., Smyth, D.J., Rowan, W.C., Nakic, N., Esparza-Gordillo, J., Bossini-Castillo, L., Tough, D.F., Larminie, C.G.C., Bronson, P.G. et al. (2019, March). Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *bioRxiv*, 566810.

Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A. et al. (2010, June). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet. 42*(6), 508–514.

Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012, February). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols 7*(3), 500–507.

Stephens, M. and Balding, D.J. (2009, October). Bayesian statistical methods for genetic association studies. *Nature Reviews. Genetics 10*(10), 681–690.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M. et al. (2013, December). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science (New York, N.Y.) 342*(6164), 1367–1372.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. et al. (2007, October). Population genomics of human gene expression. *Nature genetics 39*(10), 1217–1224.

Stunnenberg, H.G. and Hirst, M. (2016, November). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell 167*(5), 1145–1149.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A. 102*(43), 15545–15550.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. et al. (2015, March). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine 12*(3), e1001779.

Tan, L., Xing, D., Chang, C.H., Li, H. and Xie, X.S. (2018, August). Three-dimensional genome structures of single diploid human cells. *Science 361*(6405), 924–928.

Tehranchi, A.K., Myrthil, M., Martin, T., Hie, B.L., Golan, D. and Fraser, H.B. (2016, April). Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell 165*(3), 730–741.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*(7307), 707–713.

Thaventhiran, J.E.D., Allen, H.L., Burren, O.S., Farmery, J.H.R., Staples, E., Zhang, Z., Rae, W., Greene, D., Simeoni, I., Maimaris, J. et al. (2018, December). Whole Genome Sequencing of Primary Immunodeficiency reveals a role

for common and rare variants in coding and non-coding sequences. *bioRxiv*, 499988.

The ENCODE Project Consortium (2012, September). An integrated encyclopedia of DNA elements in the human genome. *Nature 489*(7414), 57–74.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. et al. (2012, September). The accessible chromatin landscape of the human genome. *Nature 489*(7414), 75–82.

Todd, J.A., Bell, J.I. and McDevitt, H.O. (1987, October). HLA-DQ Beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature 329*(6140), 599.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. and de Laat, W. (2002, December). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell 10*(6), 1453–1465.

Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G. et al. (2011, November). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics 43*(12), 1193–1201.

Trynka, G., Westra, H.J., Slowikowski, K., Hu, X., Xu, H., Stranger, B.E., Klein, R.J., Han, B. and Raychaudhuri, S. (2015). Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet. 97*(1), 139–152.

Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S. et al. (2016, June). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell 165*(6), 1530–1545.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. (2017, July). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics 101*(1), 5–22.

Võsa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S. et al. (2018, October). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*, 447367.

Wakefield, J. (2007, August). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics 81*(2), 208–227.

Wakefield, J. (2008, June). Reporting and interpretation in genome-wide association studies. *International Journal of Epidemiology 37*(3), 641–653.

Wakefield, J. (2009, January). Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol 33*(1), 79–86.

Wallace, C., Cutler, A.J., Pontikos, N., Pekalski, M.L., Burren, O.S., Cooper, J.D., García, A.R., Ferreira, R.C., Guo, H., Walker, N.M. et al. (2015, June). Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLoS Genet 11*(6), e1005272.

Wallace, C., Rotival, M., Cooper, J.D., Rice, C.M., Yang, J.H.M., McNeill, M., Smyth, D.J., Niblett, D., Cambien, F., Cardiogenics Consortium et al. (2012, June). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics 21*(12), 2815–2824.

Wang, G., Sarkar, A., Carbonetto, P. and Stephens, M. (2018, December). A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, 501114.

Wang, K., Li, M. and Bucan, M. (2007, December). Pathway-Based Approaches for Analysis of Genomewide Association Studies. *American Journal of Human Genetics 81*(6), 1278–1283.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005, February). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews. Genetics 6*(2), 109–118.

Weiss, P.F., Xiao, R., Brandon, T.G., Pagnini, I., Wright, T.B., Beukelman, T., Morgan-DeWitt, E. and Feudtner, C. (2018, January). Comparative effectiveness of tumor necrosis factor agents and disease-modifying antirheumatic therapy in children with enthesitis-related arthritis: the first year after diagnosis. *The Journal of rheumatology 45*(1), 107–114.

Wellcome Trust Case Control Consortium (2007, June). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*(7145), 661–678.

Wellcome Trust Case Control Consortium, Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S. et al. (2012, December). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics 44*(12), 1294–1301.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P. and Andrews, S. (2015, November). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research 4*.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z. et al. (2014, November). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet. 46*(11), 1173–1186.

Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007, October). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research 17*(10), 1520–1528.

Wray, N.R., Kemper, K.E., Hayes, B.J., Goddard, M.E. and Visscher, P.M. (2019). Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics 211*(4), 1131–1141.

Wu, Y., Byrne, E.M., Zheng, Z., Kemper, K.E., Yengo, L., Mallett, A.J., Yang, J., Visscher, P.M. and Wray, N.R. (2019). Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nature Communications 10*(1), 1891.

Xiao, R. and Boehnke, M. (2009, July). Quantifying and correcting for the winner's curse in genetic association studies. *Genetic epidemiology 33*(5), 453–462.

Xing, K. and He, X. (2015, April). Reassessing the "duon" hypothesis of protein evolution. *Molecular Biology and Evolution 32*(4), 1056–1062.

Yan, J., Chen, S.A.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., Rivera, C.M., Wang, C., Ye, Z. et al. (2018, February). Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Research 28*(2), 204–220.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011, January). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet 88*(1), 76–82.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. et al. (2016, January). Ensembl 2016. *Nucleic Acids Research 44*(D1), D710–D716.

Zhang, H., Massey, D., Tremelling, M. and Parkes, M. (2008). Genetics of inflammatory bowel disease: clues to pathogenesis. *British Medical Bulletin 87*, 17–30.

Zhang, X., Joehanes, R., Chen, B.H., Huan, T., Ying, S., Munson, P.J., Johnson, A.D., Levy, D. and O'Donnell, C.J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat. Genet.*.

Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., Pourcain, B.S. et al. (2017, January). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics 33*(2), 272–279.

Zhernakova, A., van Diemen, C.C. and Wijmenga, C. (2009, January). Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews. Genetics 10*(1), 43–55.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. et al. (2016, May). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics 48*(5), 481–487.

Zuin, J., Dixon, J.R., Reijden, M.I.J.A.v.d., Ye, Z., Kolovos, P., Brouwer, R.W.W., Corput, M.P.C.v.d., Werken, H.J.G.v.d., Knoch, T.A., IJcken, W.F.J.v. et al. (2014, January). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences 111*(3), 996–1001.

Appendix A

A.1 Summary of PCHi-C datasets

Cell type	Label	Biological replicates	Unique captured read pairs	Detected interactions
Megakaryocytes	MK	4	653,848,788	150,203
Erythroblasts	Ery	3	588,786,672	144,771
Neutrophils	Neu	3	736,055,569	131,609
Monocytes	Mon	3	572,357,387	151,389
Macrophages M0	Mϕ0	3	668,675,248	163,791
Macrophages M1	Mϕ1	3	497,683,496	163,399
Macrophages M2	Mϕ2	3	523,561,551	173,449
Endothelial Precursors	EndP	3	420,536,621	141,382
Naive B cells	nB	3	629,928,642	171,439
Total B cells	tB	3	702,533,922	183,119
Fetal Thymus	FetT	3	776,491,344	145,577
Naive $CD4^+$ T cells	nCD4	4	844,697,853	192,048
Total $CD4^+$ T cells	tCD4	3	836,974,777	166,668
Non-Activated Total $CD4^+$ T cells	naCD4	3	721,030,702	177,371
Activated Total $CD4^+$ T cells	aCD4	3	749,720,649	188,714
Naive $CD8^+$ T cells	nCD8	3	747,834,572	187,399
Total $CD8^+$ T cells	tCD8	3	628,771,947	183,964
Total			11,299,489,740	708,007

Table A.1 Summary of PCHi-C datasets used in this study. Adapted from Javierre et al. (2016). 'Total detected interactions' represents unique interactions captured in at least one cell type.

A.2 GWAS study references

Trait	Label	Reference
Multiple sclerosis	MS	IMSGC et al. (2011)
Celiac disease	CEL	Dubois et al. (2010)
Type 1 diabetes	T1D	Barrett et al. (2009)
Crohn's disease	CRO	Franke et al. (2010)
Primary Billiary Cirrhosis	PBC	Cordell et al. (2015)
Ulcerative colitis	UC	Anderson et al. (2011)
Systemic Lupus Erythrematosus	SLE	Bentham et al. (2015)
Rheumatoid arthritis	RA	Okada et al. (2014)
Type 2 diabetes	T2D	Morris et al. (2012)
Haemoglobin	HB	Soranzo et al. (2009)
Mean corp. haemoglobin	MCH	Soranzo et al. (2009)
Packed cell volume	PCV	Soranzo et al. (2009)
Mean corp. haemoglobin conc.	MCHC	Soranzo et al. (2009)
Red blood cell count	RBC	Soranzo et al. (2009)
Mean corpuscular volume	MCV	Soranzo et al. (2009)
Platelet count	PLT	Gieger et al. (2011)
Platelet volume	PV	Gieger et al. (2011)
Body Mass Index	BMI	Locke et al. (2015)
Low density lipoprotein	LDL	Teslovich et al. (2010)
Tryglycerides	TG	Teslovich et al. (2010)
High density lipoprotein	HDL	Teslovich et al. (2010)
Total Cholesterol	TC	Teslovich et al. (2010)
Insulin sensitivity	INS	Manning et al. (2012)
Insulin sensitivity BMI adj.	INS BMI	Manning et al. (2012)
Glucose sensitivity	GLUCOSE	Manning et al. (2012)
Glucose sensitivity BMI adj.	GLUCOSE BMI	Manning et al. (2012)
Height	HEIGHT	Wood et al. (2014)
Diastolic blood pressure	BP DIA	ICBP et al. (2011)
Systolic blood pressure	BP SYS	ICBP et al. (2011)
Lumbar spine bone mineral dens.	LSBMD	Estrada et al. (2012)
Femoral neck bone mineral dens.	FNBMD	Estrada et al. (2012)

Table A.2 Table of GWAS studies used in Chapters 2 and 3

Appendix B

B.1 COGS prioritised genes Peters et al. (2016)

Table B.1 Genes significantly (FDR<5%) differentially expressed between UC/CD and healthy controls with PCHi-C COGS scores > 0.5. Genes are ordered by COGS score.

Gene Name	COGS score	DE p_{adj}	Disease
FAIM3	1.000	0.000	UC
COX4I1	1.000	0.036	UC
RPS24	1.000	0.011	CD
IKZF1	0.999	0.001	CD
ACSS1	0.999	0.017	UC
CHD1	0.988	0.032	CD
CD274	0.987	0.002	CD
ROPN1L	0.944	0.042	UC
ADO	0.941	0.006	CD
TFAM	0.939	0.001	CD
ETS1	0.934	0.001	UC
MIDN	0.927	0.011	CD
SBNO2	0.926	0.000	CD
IPMK	0.922	0.028	CD
RQCD1	0.914	0.038	UC
STK32B	0.894	0.022	UC
CTDSP1	0.877	0.008	UC
ADAM10	0.854	0.006	UC
MYC	0.836	0.034	CD
FCGR2A	0.836	0.021	UC
FCRLA	0.835	0.018	UC
SGMS1	0.835	0.000	UC
CD244	0.823	0.003	CD
PIM3	0.822	0.001	UC
BCL6	0.811	0.001	UC
RTP2	0.800	0.043	UC
RASGRP1	0.795	0.005	CD
IKZF3	0.793	0.000	UC
LYRM7	0.771	0.001	UC
DGKD	0.758	0.004	CD
CHRNE	0.737	0.041	UC
ARRB2	0.717	0.001	UC
PIP4K2C	0.686	0.009	UC

Continued on next page

Gene Name	COGS score	DE p_{adj}	Disease
ADAM9	0.684	0.000	UC
GATA3	0.679	0.031	UC
PAPD7	0.659	0.007	UC
MFF	0.658	0.020	UC
IGF2	0.649	0.049	UC
STK36	0.634	0.013	UC
GPX4	0.631	0.000	CD
BEST1	0.627	0.001	CD
ATG4D	0.626	0.017	CD
AGAP2	0.621	0.017	UC
SMAD3	0.612	0.013	CD
MRPL4	0.599	0.017	CD
MARCH9	0.594	0.003	UC
MAN2A2	0.591	0.038	CD
GALC	0.589	0.001	CD
CDK4	0.588	0.014	UC
MLC1	0.575	0.031	CD
DSE	0.573	0.000	UC
BCL2	0.565	0.015	UC
FBL	0.561	0.001	UC
C9orf37	0.543	0.001	UC
DPP7	0.542	0.047	UC
SSNA1	0.542	0.000	UC
PHPT1	0.542	0.007	UC
CCDC183	0.542	0.012	UC
ABCA2	0.542	0.021	UC
SCAMP3	0.529	0.020	CD
TTC1	0.523	0.045	UC
ZBTB49	0.520	0.047	UC
CD55	0.517	0.002	UC
FYB	0.508	0.008	CD
ATF6	0.504	0.003	UC

B.2 Prioritised COGS genes from Burren et al. (2017)

Table B.2 Activated and non-activated $CD4^+$ T cell PCHi-C COGS prioritised genes (COGS > 0.5) across 6 IMD studies. All diseases were based on ImmunoChip studies (IC) apart from SLE which only used a genome-wide (GW) study. Note that RA prioritisation is based on both IC and GW studies. Context indicates labelling by the hierarchical COGS method where prioritisation is based on; INT = interaction, CODE = coding SNP, NCODE = interaction + promoter, PROM = promoter, NACT = Non activated $CD4^+$ T cells, ACT = Activated $CD4^+$ T cells and OVERALL = interaction + promoter + coding SNP. 'Expr' (Expression) and 'eRNA' columns reflect differential expression of target genes and enhancer RNAs between non activated $CD4^+$ T cells respectively; ND = Not detected, '+' = Up , '-' = Down , '=' = No change and are relative to activation. Locus and SNP are the most proximal disease-matched susceptibility region and index SNP respectively as curated in ImmunoBase. A list of COGS scores for all genes analysed is available as a supplementary table in Burren et al. (2017).

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
1	ENSG00000160087	UBE2J2	SLE	GW	0.60	INT	+	ND			
2	ENSG00000157870	FAM213B	RA	GW	0.75	PROM	=	=	1p36.32	chr1:2523811	7.8E-14
3	ENSG00000142606	MMEL1	RA	GW	0.76	PROM	ND	=	1p36.32	chr1:2523811	7.8E-14
4	ENSG00000134242	PTPN22	ATD	IC	0.57	CODE	+	ND			
5	ENSG00000134242	PTPN22	RA	IC	0.66	CODE	+	ND			
6	ENSG00000134242	PTPN22	RA	GW	0.68	CODE	+	ND			
7	ENSG00000134242	PTPN22	T1D	IC	0.82	CODE	+	ND			
8	ENSG00000134247	PTGFRN	RA	GW	0.98	INT	+	ND	1p13.2	rs2476601	6.6E-170
9	ENSG00000134256	CD101	RA	GW	0.98	INT	-	ND	1p13.2	rs2476601	6.6E-170
10	ENSG00000116830	TTF2	RA	GW	0.97	INT	+	ND	1p13.2	rs2476601	6.6E-170
11	ENSG00000134253	TRIM45	RA	GW	0.97	INT	ND	ND	1p13.2	rs2476601	6.6E-170
12	ENSG00000160712	IL6R	RA	GW	0.56	NACT	-	ND	1q21.3	rs2228145	3.7E-09

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
13	ENSG00000162706	CADM3	SLE	GW	0.90	CODE	ND	ND			
14	ENSG00000143226	FCGR2A	RA	GW	0.52	PROM	ND	ND			
15	ENSG00000143226	FCGR2A	RA	IC	0.59	PROM	ND	ND			
16	ENSG00000143226	FCGR2A	SLE	GW	0.96	PROM	ND	ND	1q23.3	rs1801274	1.0E-12
17	ENSG00000132185	FCRLA	RA	GW	0.60	INT	=	ND			
18	ENSG00000132185	FCRLA	RA	IC	0.69	INT	=	ND			
19	ENSG00000132185	FCRLA	SLE	GW	0.89	INT	=	ND	1q23.3	rs1801274	1.0E-12
20	ENSG00000162746	FCRLB	RA	GW	0.61	INT	ND	ND			
21	ENSG00000162746	FCRLB	RA	IC	0.69	INT	ND	ND			
22	ENSG00000162746	FCRLB	SLE	GW	0.89	INT	ND	ND	1q23.3	rs1801274	1.0E-12
23	ENSG00000081721	DUSP12	RA	GW	0.60	ACT	=	ND			
24	ENSG00000081721	DUSP12	RA	IC	0.69	INT	=	ND			
25	ENSG00000081721	DUSP12	SLE	GW	0.89	ACT	=	ND	1q23.3	rs1801274	1.0E-12
26	ENSG00000198821	CD247	RA	GW	0.53	PROM	-	+			
27	ENSG00000116701	NCF2	SLE	GW	1.00	CODE	-	ND			
28	ENSG00000116750	UCHL5	CEL	IC	1.00	ACT	=	ND	1q31.2	rs2816316	6.7E-25
29	ENSG00000116747	TROVE2	CEL	IC	1.00	ACT	+	ND	1q31.2	rs2816316	6.7E-25
30	ENSG00000136634	IL10	T1D	IC	0.67	PROM	+	+	1q32.1	rs3024493	2.0E-08
31	ENSG00000142224	IL19	T1D	IC	1.00	INT	ND	+	1q32.1	rs3024493	2.0E-08
32	ENSG00000162891	IL20	T1D	IC	0.67	INT	ND	+	1q32.1	rs3024493	2.0E-08
33	ENSG00000162892	IL24	T1D	IC	1.00	INT	+	+	1q32.1	rs3024493	2.0E-08

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
34	ENSG00000162894	FAIM3	T1D	IC	1.00	INT	-	+	1q32.1	rs3024493	2.0E-08
35	ENSG00000162896	PIGR	T1D	IC	1.00	INT	ND	+	1q32.1	rs3024493	2.0E-08
36	ENSG00000162897	FCAMR	T1D	IC	0.67	INT	ND	+	1q32.1	rs3024493	2.0E-08
37	ENSG00000152518	ZFP36L2	CEL	IC	0.82	INT	-	-			
38	ENSG00000115421	PAPOLG	CEL	IC	1.00	INT	=	ND	2p16.1	rs13003464	4.3E-16
39	ENSG00000162924	REL	RA	IC	0.59	NCODE	+	ND	2p16.1	rs34695944	1.7E-15
40	ENSG00000162924	REL	RA	GW	0.93	NCODE	+	ND	2p16.1	rs34695944	1.7E-15
41	ENSG00000162924	REL	CEL	IC	1.00	INT	+	ND	2p16.1	rs13003464	4.3E-16
42	ENSG00000233404	FLJ20373	CEL	IC	0.77	INT	ND	ND	2q12.1	rs990171	1.2E-16
43	ENSG00000115590	IL1R2	CEL	IC	0.77	INT	+	ND	2q12.1	rs990171	1.2E-16
44	ENSG00000115598	IL1RL2	CEL	IC	0.62	INT	+	ND	2q12.1	rs990171	1.2E-16
45	ENSG00000152127	MGAT5	SLE	GW	0.99	INT	+	ND			
46	ENSG00000115232	ITGA4	CEL	IC	0.99	INT	-	ND	2q31.3	rs1018326	3.1E-16
47	ENSG00000151690	MFSD6	RA	GW	0.86	INT	=	ND	2q32.3	rs13426947	7.2E-10
48	ENSG00000189362	TMEM194B	RA	GW	0.95	ACT	+	ND	2q32.3	rs13426947	7.2E-10
49	ENSG00000138386	NAB1	SLE	GW	0.72	CODE	+	ND			
50	ENSG00000138378	STAT4	RA	GW	0.52	INT	+	ND	2q32.3	rs13426947	7.2E-10
51	ENSG00000138378	STAT4	RA	IC	0.53	INT	+	ND	2q32.3	rs13426947	7.2E-10
52	ENSG00000138378	STAT4	CEL	IC	0.99	INT	+	ND	2q32.3	rs6715106	8.4E-09
53	ENSG00000173559	NABP1	RA	IC	0.58	INT	+	ND	2q32.3	rs13426947	7.2E-10
54	ENSG00000173559	NABP1	RA	GW	0.60	INT	+	ND	2q32.3	rs13426947	7.2E-10

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
55	ENSG00000173559	NABP1	CEL	IC	0.99	INT	+	ND	2q32.3	rs6715106	8.4E-09
56	ENSG00000119004	CYP20A1	T1D	IC	0.66	INT	+	+	2q33.2	rs3087243	7.4E-21
57	ENSG00000173166	RAPH1	T1D	IC	0.83	INT	+	+	2q33.2	rs3087243	7.4E-21
58	ENSG00000178562	CD28	T1D	IC	0.81	INT	=	+	2q33.2	rs3087243	7.4E-21
59	ENSG00000187118	CMC1	RA	GW	0.94	INT	=	ND	3p24.1	rs3806624	8.6E-09
60	ENSG00000163512	AZI2	RA	GW	0.56	INT	-	ND	3p24.1	rs3806624	8.6E-09
61	ENSG00000206559	ZCWPW2	RA	GW	0.56	INT	ND	ND	3p24.1	rs3806624	8.6E-09
62	ENSG00000136068	FLNB	RA	GW	0.50	NACT	=	ND	3p14.3	rs35677470	1.7E-07
63	ENSG00000163687	DNASE1L3	RA	GW	0.65	CODE	ND	ND			
64	ENSG00000163687	DNASE1L3	RA	IC	0.82	CODE	ND	ND			
65	ENSG00000114850	SSR3	RA	GW	0.80	INT	=	ND			
66	ENSG00000163659	TIPARP	RA	GW	0.78	INT	-	ND			
67	ENSG00000197980	LEKR1	RA	GW	0.75	INT	=	ND			
68	ENSG00000213186	TRIM59	CEL	IC	1.00	NACT	=	ND	3q25.33	rs1353248	9.8E-09
69	ENSG00000114209	PDCD10	SLE	GW	0.54	INT	=	ND			
70	ENSG00000163536	SERPINI1	SLE	GW	0.54	INT	-	ND			
71	ENSG00000145495	MARCH6	RA	GW	0.61	INT	=	ND			
72	ENSG00000145491	ROPN1L	RA	GW	0.83	INT	ND	ND			
73	ENSG00000154122	ANKH	RA	GW	0.67	INT	-	ND			
74	ENSG00000134352	IL6ST	RA	GW	1.00	INT	+	ND	5q11.2	rs7731626	7.3E-24
75	ENSG00000134352	IL6ST	RA	IC	1.00	INT	+	ND	5q11.2	rs6859219	4.0E-16

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
76	ENSG00000153922	CHD1	RA	IC	0.62	INT	+	ND	5q21.1	rs39984	9.3E-08
77	ENSG00000153922	CHD1	RA	GW	0.64	INT	+	ND	5q21.1	rs39984	9.3E-08
78	ENSG00000153922	CHD1	SLE	GW	0.67	INT	+	ND			
79	ENSG00000174132	FAM174A	SLE	GW	0.62	INT	-	ND			
80	ENSG00000113532	ST8SIA4	SLE	GW	0.53	INT	+	ND			
81	ENSG00000145723	GIN1	RA	GW	0.61	INT	-	ND	5q21.1	rs39984	9.3E-08
82	ENSG00000145723	GIN1	RA	IC	0.62	INT	-	ND	5q21.1	rs39984	9.3E-08
83	ENSG00000145725	PPIP5K2	RA	GW	0.61	INT	=	ND	5q21.1	rs39984	9.3E-08
84	ENSG00000145725	PPIP5K2	RA	IC	0.62	INT	=	ND	5q21.1	rs39984	9.3E-08
85	ENSG00000113552	GNPDA1	CEL	IC	0.90	INT	+	+			
86	ENSG00000197043	ANXA6	RA	GW	0.76	INT	-	ND			
87	ENSG00000197043	ANXA6	SLE	GW	1.00	INT	-	ND	5q33.1	rs7708392	3.8E-13
88	ENSG00000113328	CCNG1	SLE	GW	1.00	INT	-	ND	5q33.3	rs2431697	8.0E-28
89	ENSG00000170584	NUDCD2	SLE	GW	1.00	INT	+	ND	5q33.3	rs2431697	8.0E-28
90	ENSG00000072571	HMMR	SLE	GW	1.00	INT	-	ND	5q33.3	rs2431697	8.0E-28
91	ENSG00000137265	IRF4	CEL	IC	0.95	CODE	+	ND			
92	ENSG00000157593	SLC35B2	RA	GW	0.56	PROM	+	ND	6p21.1	rs2233424	1.4E-19
93	ENSG00000146232	NFKBIE	RA	GW	0.76	OVERALL	+	ND	6p21.1	rs2233424	1.4E-19
94	ENSG00000178233	TMEM151B	RA	GW	0.60	PROM	ND	ND	6p21.1	rs2233424	1.4E-19
95	ENSG00000272442	RP11-444E17.6	RA	GW	0.58	PROM	ND	ND	6p21.1	rs2233424	1.4E-19

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
96	ENSG00000112159	MDN1	RA	IC	0.53	INT	+	+	6q15	rs72928038	8.2E-07
97	ENSG00000112159	MDN1	ATD	IC	0.88	INT	+	+	6q15	rs72928038	1.2E-07
98	ENSG00000112159	MDN1	T1D	IC	0.99	INT	+	+	6q15	rs72928038	6.4E-14
99	ENSG00000057657	PRDM1	RA	GW	0.55	INT	+	ND	6q21	rs9372120	7.6E-10
100	ENSG00000118503	TNFAIP3	RA	GW	0.79	ACT	+	ND	6q23.3	rs17264332	3.9E-20
101	ENSG00000118503	TNFAIP3	RA	IC	0.83	ACT	+	ND	6q23.3	rs17264332	3.9E-20
102	ENSG00000118503	TNFAIP3	CEL	IC	0.98	ACT	+	ND	6q23.3	rs17264332	5.0E-30
103	ENSG00000146425	DYNLT1	RA	GW	0.98	ACT	=	+	6q25.3	rs2451258	2.7E-11
104	ENSG00000146425	DYNLT1	CEL	IC	0.75	INT	=	ND	6q25.3	rs182429	8.5E-16
105	ENSG00000164674	SYTL3	RA	GW	0.98	ACT	+	+	6q25.3	rs2451258	2.7E-11
106	ENSG00000164674	SYTL3	CEL	IC	0.75	INT	+	ND	6q25.3	rs182429	8.5E-16
107	ENSG00000203711	C6orf99	CEL	IC	0.74	INT	ND	ND	6q25.3	rs182429	8.5E-16
108	ENSG00000164691	TAGAP	RA	GW	0.98	INT	+	+	6q25.3	rs2451258	2.7E-11
109	ENSG00000164691	TAGAP	CEL	IC	1.00	OVERALL	+	ND	6q25.3	rs182429	8.5E-16
110	ENSG00000060762	MPC1	ATD	IC	0.53	INT	-	ND			
111	ENSG00000249141	RP11-514O12.4	ATD	IC	0.58	NCODE	ND	ND			
112	ENSG00000026297	RNASET2	ATD	IC	0.67	NCODE	=	ND			
113	ENSG00000213066	FGFR1OP	RA	GW	1.00	NACT	=	ND	6q27	rs1571878	5.0E-35
114	ENSG00000106546	AHR	RA	GW	0.96	ACT	+	ND			
115	ENSG00000155849	ELMO1	RA	IC	0.67	NCODE	+	ND	7p14.1	rs79758729	9.2E-07

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
116	ENSG00000155849	ELMO1	CEL	IC	0.74	NCODE	+	ND	7p14.1	1kg_7_37384979	2.1E-08
117	ENSG00000136273	HUS1	SLE	GW	0.66	INT	+	ND	7p12.2	rs4917014	6.4E-14
118	ENSG00000042813	ZPBP	T1D	IC	0.91	INT	ND	ND	7p12.2	rs10272724	4.8E-09
119	ENSG00000164500	C7orf72	T1D	IC	0.91	INT	ND	ND	7p12.2	rs10272724	4.8E-09
120	ENSG00000185811	IKZF1	SLE	GW	0.74	INT	=	ND	7p12.2	rs4917014	6.4E-14
121	ENSG00000185811	IKZF1	T1D	IC	0.99	INT	=	ND	7p12.2	rs10272724	4.8E-09
122	ENSG00000132436	FIGNL1	SLE	GW	0.70	INT	+	ND	7p12.2	rs4917014	6.4E-14
123	ENSG00000105810	CDK6	RA	GW	0.96	OVERALL	+	ND	7q21.2	rs4272	5.0E-09
124	ENSG00000147316	MCPH1	SLE	GW	0.51	INT	+	ND	8p23.1	rs2736340	6.3E-20
125	ENSG00000155189	AGPAT5	SLE	GW	0.58	INT	+	ND	8p23.1	rs2736340	6.3E-20
126	ENSG00000147324	MFHAS1	SLE	GW	0.66	INT	=	ND	8p23.1	rs2736340	6.3E-20
127	ENSG00000171056	SOX7	SLE	GW	0.54	ACT	ND	ND	8p23.1	rs2736340	6.3E-20
128	ENSG00000258724	SOX7	SLE	GW	0.54	ACT	ND	ND	8p23.1	rs2736340	6.3E-20
129	ENSG00000254093	PINX1	SLE	GW	0.54	ACT	+	ND	8p23.1	rs2736340	6.3E-20
130	ENSG00000154319	FAM167A	SLE	GW	0.55	PROM	ND	ND	8p23.1	rs2736340	6.3E-20
131	ENSG00000154319	FAM167A	RA	GW	0.56	PROM	ND	ND	8p23.1	rs2736337	3.1E-13
132	ENSG00000136573	BLK	RA	GW	0.52	PROM	-	ND	8p23.1	rs2736337	3.1E-13
133	ENSG00000136573	BLK	SLE	GW	0.53	PROM	-	ND	8p23.1	rs2736340	6.3E-20
134	ENSG00000255394	C8orf49	SLE	GW	0.76	INT	ND	ND	8p23.1	rs2736340	6.3E-20
135	ENSG00000154328	NEIL2	SLE	GW	0.76	INT	=	ND	8p23.1	rs2736340	6.3E-20
136	ENSG00000079459	FDFT1	SLE	GW	0.54	INT	=	ND	8p23.1	rs2736340	6.3E-20

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
137	ENSG00000255046	RP11-297N6.4	SLE	GW	0.54	INT	=	ND	8p23.1	rs2736340	6.3E-20
138	ENSG00000164733	CTSB	SLE	GW	0.54	INT	-	ND	8p23.1	rs2736340	6.3E-20
139	ENSG00000205883	DEFB135	SLE	GW	0.57	NACT	ND	ND	8p23.1	rs2736340	6.3E-20
140	ENSG00000136997	MYC	RA	GW	0.70	INT	+	+	8q24.21	rs1516971	1.3E-10
141	ENSG00000095261	PSMD5	RA	GW	0.73	INT	=	+	9q33.2	rs10985070	5.0E-11
142	ENSG00000056558	TRAF1	RA	GW	0.89	NCODE	+	+	9q33.2	rs10985070	5.0E-11
143	ENSG00000119397	CNTRL	RA	GW	0.85	INT	-	+	9q33.2	rs10985070	5.0E-11
144	ENSG00000134460	IL2RA	RA	GW	1.00	PROM	+	+	10p15.1	rs706778	4.6E-14
145	ENSG00000134460	IL2RA	RA	IC	1.00	PROM	+	+	10p15.1	rs706778	4.6E-14
146	ENSG00000134460	IL2RA	ATD	IC	1.00	NCODE	+	+	10p15.1	rs706779	2.7E-07
147	ENSG00000134460	IL2RA	T1D	IC	1.00	NCODE	+	ND	10p15.1	rs41295121	4.9E-08
148	ENSG00000134453	RBM17	RA	IC	0.67	ACT	+	+	10p15.1	rs706778	4.6E-14
149	ENSG00000134453	RBM17	RA	GW	0.77	ACT	+	+	10p15.1	rs706778	4.6E-14
150	ENSG00000134453	RBM17	ATD	IC	0.99	INT	+	+	10p15.1	rs706779	2.7E-07
151	ENSG00000134453	RBM17	T1D	IC	1.00	NCODE	+	ND	10p15.1	rs41295121	4.9E-08
152	ENSG00000170525	PFKFB3	RA	GW	0.65	NCODE	+	ND	10p15.1	rs10795791	3.0E-06
153	ENSG00000212743	DKFZP667-F0711	RA	GW	0.75	PROM	=	+	10p15.1	rs947474	4.1E-10
154	ENSG00000212743	DKFZP667-F0711	CEL	IC	0.76	OVERALL	=	+	10p15.1	rs2387397	1.9E-08

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
155	ENSG00000212743	DKFZP667-F0711	T1D	IC	0.86	NACT	=	ND	10p15.1	rs41295121	4.9E-08
156	ENSG00000065675	PRKCQ	ATD	IC	0.54	INT	-	+	10p15.1	rs706779	2.7E-07
157	ENSG00000107485	GATA3	RA	GW	1.00	PROM	=	=	10p14	rs3824660	2.0E-08
158	ENSG00000107485	GATA3	SLE	GW	0.68	INT	=	ND			
159	ENSG00000128815	WDFY4	SLE	GW	0.88	CODE	=	ND			
160	ENSG00000150347	ARID5B	RA	GW	0.69	INT	+	+	10q21.2	rs12764378	4.5E-10
161	ENSG00000150347	ARID5B	SLE	GW	0.99	INT	+	+	10q21.2	rs4948496	1.0E-10
162	ENSG00000185507	IRF7	SLE	GW	0.69	NCODE	=	ND	11p15.5	rs4963128	4.9E-09
163	ENSG00000099834	CDHR5	SLE	GW	0.84	OVERALL	ND	ND	11p15.5	rs4963128	4.9E-09
164	ENSG00000185940	KRTAP5-5	T1D	IC	1.00	INT	ND	ND	11p15.5	rs72853903	6.2E-10
165	ENSG00000205864	KRTAP5-6	T1D	IC	1.00	INT	ND	ND	11p15.5	rs72853903	6.2E-10
166	ENSG00000129965	INS-IGF2	T1D	IC	1.00	PROM	ND	ND	11p15.5	rs72853903	6.2E-10
167	ENSG00000254647	INS	T1D	IC	1.00	PROM	ND	ND	11p15.5	rs72853903	6.2E-10
168	ENSG00000180176	TH	T1D	IC	1.00	PROM	ND	ND	11p15.5	rs72853903	6.2E-10
169	ENSG00000184281	TSSC4	T1D	IC	1.00	INT	+	ND	11p15.5	rs72853903	6.2E-10
170	ENSG00000026508	CD44	SLE	GW	0.83	INT	+	+	11p13	rs2732549	1.2E-23
171	ENSG00000135362	PRR5L	RA	GW	0.68	INT	=	ND			
172	ENSG00000178105	DDX10	RA	GW	0.71	ACT	+	ND	11q22.3	chr11:107967350	1.4E-08

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
173	ENSG00000110367	DDX6	CEL	IC	0.51	INT	=	ND	11q23.3	rs10892258	1.7E-11
174	ENSG00000134954	ETS1	CEL	IC	1.00	INT	-	=	11q24.3	rs11221332	5.3E-16
175	ENSG00000134954	ETS1	SLE	GW	0.80	ACT	-	ND	11q24.3	rs7941765	1.3E-10
176	ENSG00000134954	ETS1	RA	GW	0.99	ACT	-	ND	11q24.3	rs73013527	1.2E-10
177	ENSG00000151702	FLI1	CEL	IC	0.71	INT	-	=	11q24.3	rs11221332	5.3E-16
178	ENSG00000118971	CCND2	SLE	GW	0.79	INT	+	ND			
179	ENSG00000139531	SUOX	RA	GW	0.62	PROM	=	ND	12q13.2	rs773125	1.1E-10
180	ENSG00000065361	ERBB3	T1D	IC	0.65	PROM	ND	ND	12q13.2	rs11171739	9.7E-11
181	ENSG00000135506	OS9	RA	GW	0.57	PROM	-	ND	12q13.2	rs773125	1.1E-10
182	ENSG00000135439	AGAP2	RA	GW	0.51	INT	-	ND	12q13.2	rs773125	1.1E-10
183	ENSG00000135452	TSPAN31	RA	GW	0.51	INT	+	ND	12q13.2	rs773125	1.1E-10
184	ENSG00000111252	SH2B3	CEL	IC	0.70	CODE	+	ND			
185	ENSG00000111252	SH2B3	T1D	IC	0.84	CODE	+	ND			
186	ENSG00000134882	UBAC2	T1D	IC	0.86	INT	-	+	13q32.3	rs9585056	5.2E-09
187	ENSG00000134882	UBAC2	CEL	IC	0.67	INT	-	-			
188	ENSG00000125245	GPR18	T1D	IC	0.86	INT	-	+	13q32.3	rs9585056	5.2E-09
189	ENSG00000125245	GPR18	CEL	IC	0.67	INT	-	-			
190	ENSG00000169508	GPR183	T1D	IC	0.86	INT	+	+	13q32.3	rs9585056	5.2E-09
191	ENSG00000169508	GPR183	CEL	IC	0.67	INT	+	-			
192	ENSG00000185650	ZFP36L1	RA	GW	0.97	INT	+	ND	14q24.1	rs1950897	8.2E-11
193	ENSG00000165409	TSHR	ATD	IC	1.00	INT	ND	ND	14q31.1	rs2300519	1.3E-38

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
194	ENSG00000165417	GTF2A1	ATD	IC	1.00	INT	=	ND	14q31.1	rs2300519	1.3E-38
195	ENSG00000197406	DIO3	T1D	IC	0.76	INT	ND	ND	14q32.2	rs56994090	1.1E-11
196	ENSG00000172575	RASGRP1	RA	IC	0.96	INT	=	+	15q14	rs8043085	1.4E-10
197	ENSG00000172575	RASGRP1	RA	GW	1.00	NCODE	=	+	15q14	rs8032939	1.9E-18
198	ENSG00000172575	RASGRP1	T1D	IC	0.91	INT	=	ND	15q14	rs12908309	4.3E-08
199	ENSG00000175779	C15orf53	T1D	IC	0.81	INT	ND	ND	15q14	rs12908309	4.3E-08
200	ENSG00000175779	C15orf53	RA	IC	0.96	INT	ND	+	15q14	rs8043085	1.4E-10
201	ENSG00000175779	C15orf53	RA	GW	1.00	INT	ND	+	15q14	rs8032939	1.9E-18
202	ENSG00000140332	TLE3	RA	GW	0.82	INT	+	ND	15q23	rs8026898	3.6E-19
203	ENSG00000182108	DEXI	T1D	IC	0.51	INT	=	ND	16p13.13	rs193778	4.4E-10
204	ENSG00000182108	DEXI	RA	GW	0.85	INT	=	ND	16p13.13	rs4780401	4.1E-08
205	ENSG00000038532	CLEC16A	T1D	IC	0.55	INT	=	ND	16p13.13	rs193778	4.4E-10
206	ENSG00000038532	CLEC16A	RA	GW	0.90	INT	=	ND	16p13.13	rs4780401	4.1E-08
207	ENSG00000175643	RMI2	SLE	GW	0.60	INT	-	ND	16p13.13	rs9652601	7.4E-17
208	ENSG00000175643	RMI2	T1D	IC	0.97	NCODE	-	ND	16p13.13	rs193778	4.4E-10
209	ENSG00000185338	SOCS1	T1D	IC	0.85	NCODE	+	ND	16p13.13	rs12927355	3.0E-22
210	ENSG00000178279	TNP2	T1D	IC	0.59	INT	ND	ND	16p13.13	rs12927355	3.0E-22
211	ENSG00000178257	PRM3	T1D	IC	0.59	INT	ND	ND	16p13.13	rs12927355	3.0E-22
212	ENSG00000122304	PRM2	SLE	GW	0.52	ACT	ND	ND	16p13.13	rs9652601	7.4E-17
213	ENSG00000122304	PRM2	RA	GW	0.62	ACT	ND	ND	16p13.13	rs4780401	4.1E-08
214	ENSG00000233232	NPIPB7	T1D	IC	0.66	PROM	ND	ND	16p11.2	rs151234	4.8E-11

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
215	ENSG00000261832	CLN3	T1D	IC	0.66	PROM	ND	ND	16p11.2	rs151234	4.8E-11
216	ENSG00000188603	CLN3	T1D	IC	0.66	PROM	-	ND	16p11.2	rs151234	4.8E-11
217	ENSG00000184730	APOBR	T1D	IC	0.99	OVERALL	-	ND	16p11.2	rs151234	4.8E-11
218	ENSG00000176046	NUPR1	T1D	IC	0.66	NACT	ND	ND	16p11.2	rs151234	4.8E-11
219	ENSG00000157423	HYDIN	SLE	GW	0.64	INT	ND	ND			
220	ENSG00000103091	WDR59	T1D	IC	0.95	INT	=	ND	16q23.1	rs8056814	3.0E-19
221	ENSG00000186187	ZNRF1	T1D	IC	0.95	INT	+	ND	16q23.1	rs8056814	3.0E-19
222	ENSG00000168928	CTRB2	T1D	IC	1.00	PROM	ND	ND	16q23.1	rs8056814	3.0E-19
223	ENSG00000168925	CTRB1	T1D	IC	1.00	PROM	ND	ND	16q23.1	rs8056814	3.0E-19
224	ENSG00000131148	EMC8	SLE	GW	0.97	ACT	+	ND	16q24.1	rs11644034	9.6E-18
225	ENSG00000131148	EMC8	RA	GW	1.00	ACT	+	ND	16q24.1	rs13330176	1.4E-12
226	ENSG00000131143	COX4I1	SLE	GW	0.97	ACT	=	ND	16q24.1	rs11644034	9.6E-18
227	ENSG00000131143	COX4I1	RA	GW	1.00	ACT	=	ND	16q24.1	rs13330176	1.4E-12
228	ENSG00000140968	IRF8	SLE	GW	0.99	ACT	+	ND	16q24.1	rs11644034	9.6E-18
229	ENSG00000140968	IRF8	RA	GW	1.00	ACT	+	ND	16q24.1	rs13330176	1.4E-12
230	ENSG00000174327	SLC16A13	SLE	GW	0.98	ACT	ND	ND	17p13.2	rs2286672	2.9E-09
231	ENSG00000132522	GPS2	SLE	GW	0.98	ACT	=	ND	17p13.2	rs2286672	2.9E-09
232	ENSG00000261915	RP11-542C16.2	SLE	GW	0.98	ACT	ND	ND	17p13.2	rs2286672	2.9E-09
233	ENSG00000215041	NEURL4	SLE	GW	0.99	PROM	=	ND	17p13.2	rs2286672	2.9E-09
234	ENSG00000072818	ACAP1	SLE	GW	0.99	PROM	-	ND	17p13.2	rs2286672	2.9E-09

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
235	ENSG00000170175	CHRNB1	SLE	GW	0.97	ACT	-	ND	17p13.2	rs2286672	2.9E-09
236	ENSG00000174282	ZBTB4	SLE	GW	0.98	INT	-	ND	17p13.2	rs2286672	2.9E-09
237	ENSG00000259224	SLC35G6	SLE	GW	0.98	INT	ND	ND	17p13.2	rs2286672	2.9E-09
238	ENSG00000181222	POLR2A	SLE	GW	0.98	INT	=	ND	17p13.2	rs2286672	2.9E-09
239	ENSG00000239697	TNFSF12	SLE	GW	0.98	INT	=	ND	17p13.2	rs2286672	2.9E-09
240	ENSG00000248871	TNFSF12-TNFSF13	SLE	GW	0.98	INT	ND	ND	17p13.2	rs2286672	2.9E-09
241	ENSG00000161955	TNFSF13	SLE	GW	0.98	INT	ND	ND	17p13.2	rs2286672	2.9E-09
242	ENSG00000161956	SENP3	SLE	GW	0.98	INT	+	ND	17p13.2	rs2286672	2.9E-09
243	ENSG00000161960	EIF4A1	SLE	GW	0.99	INT	+	ND	17p13.2	rs2286672	2.9E-09
244	ENSG00000129226	CD68	SLE	GW	0.99	INT	ND	ND	17p13.2	rs2286672	2.9E-09
245	ENSG00000129255	MPDU1	SLE	GW	0.99	INT	+	ND	17p13.2	rs2286672	2.9E-09
246	ENSG00000129194	SOX15	SLE	GW	0.99	INT	ND	ND	17p13.2	rs2286672	2.9E-09
247	ENSG00000129245	FXR2	SLE	GW	0.98	ACT	=	ND	17p13.2	rs2286672	2.9E-09
248	ENSG00000268503	AC007421.1	SLE	GW	0.98	ACT	ND	ND	17p13.2	rs2286672	2.9E-09
249	ENSG00000129214	SHBG	SLE	GW	0.98	ACT	ND	ND	17p13.2	rs2286672	2.9E-09
250	ENSG00000141510	TP53	SLE	GW	0.98	ACT	=	ND	17p13.2	rs2286672	2.9E-09
251	ENSG00000141499	WRAP53	SLE	GW	0.98	INT	-	ND	17p13.2	rs2286672	2.9E-09
252	ENSG00000179148	ALOXE3	SLE	GW	0.98	INT	ND	ND	17p13.2	rs2286672	2.9E-09
253	ENSG00000179094	PER1	SLE	GW	0.97	ACT	=	ND	17p13.2	rs2286672	2.9E-09
254	ENSG00000161405	IKZF3	RA	IC	0.55	PROM	+	ND	17q12	chr17:38031857	2.3E-12

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
255	ENSG00000161405	IKZF3	RA	GW	0.66	PROM	+	ND	17q12	chr17:38031857	2.3E-12
256	ENSG00000186075	ZPBP2	T1D	IC	0.51	PROM	ND	ND	17q12	rs12453507	1.0E-08
257	ENSG00000186075	ZPBP2	RA	IC	0.60	PROM	ND	ND	17q12	chr17:38031857	2.3E-12
258	ENSG00000186075	ZPBP2	RA	GW	0.71	PROM	ND	ND	17q12	chr17:38031857	2.3E-12
259	ENSG00000120068	HOXB8	SLE	GW	0.54	INT	ND	ND			
260	ENSG00000170689	HOXB9	SLE	GW	0.52	INT	ND	ND			
261	ENSG00000184557	SOCS3	RA	GW	0.58	PROM	+	ND			
262	ENSG00000087157	PGS1	RA	GW	0.55	PROM	=	ND			
263	ENSG00000175354	PTPN2	RA	GW	0.73	ACT	+	ND	18p11.21	rs8083786	6.3E-18
264	ENSG00000175354	PTPN2	T1D	IC	0.91	ACT	+	ND	18p11.21	rs1893217	1.2E-15
265	ENSG00000099625	C19orf26	SLE	GW	0.59	INT	ND	ND			
266	ENSG00000099624	ATP5D	SLE	GW	0.59	INT	=	ND			
267	ENSG00000167470	MIDN	SLE	GW	0.59	INT	+	ND			
268	ENSG00000099622	CIRBP	SLE	GW	0.59	INT	=	ND			
269	ENSG00000267303	CTD-2369P2.12	RA	GW	0.58	CODE	ND	ND			
270	ENSG00000161847	RAVER1	RA	GW	0.58	CODE	=	ND			
271	ENSG00000105397	TYK2	RA	IC	1.00	CODE	=	ND			
272	ENSG00000105397	TYK2	T1D	IC	1.00	CODE	=	ND			
273	ENSG00000105401	CDC37	RA	GW	0.96	INT	+	ND	19p13.2	chr19:10771941	8.6E-10
274	ENSG00000180739	S1PR5	RA	GW	0.96	INT	-	ND	19p13.2	chr19:10771941	8.6E-10

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
275	ENSG00000130734	ATG4D	RA	GW	0.96	INT	-	ND	19p13.2	chr19:10771941	8.6E-10
276	ENSG00000129347	KRI1	RA	GW	0.96	INT	+	ND	19p13.2	chr19:10771941	8.6E-10
277	ENSG00000129355	CDKN2D	RA	GW	0.97	INT	+	ND	19p13.2	chr19:10771941	8.6E-10
278	ENSG00000129353	SLC44A2	RA	GW	0.96	INT	-	ND	19p13.2	chr19:10771941	8.6E-10
279	ENSG00000129351	ILF3	RA	GW	0.96	PROM	+	ND	19p13.2	chr19:10771941	8.6E-10
280	ENSG00000105063	PPP6R1	SLE	GW	0.79	PROM	=	ND			
281	ENSG00000133265	HSPBP1	SLE	GW	0.79	PROM	+	ND			
282	ENSG00000160469	BRSK1	SLE	GW	0.91	PROM	ND	ND			
283	ENSG00000133247	SUV420H2	SLE	GW	0.78	ACT	-	ND			
284	ENSG00000267531	AC020922.1	SLE	GW	0.78	ACT	ND	ND			
285	ENSG00000233493	TMEM238	SLE	GW	0.92	INT	=	ND			
286	ENSG00000108107	RPL28	SLE	GW	0.93	INT	=	ND			
287	ENSG00000108106	UBE2S	SLE	GW	0.92	INT	+	ND			
288	ENSG00000187902	SHISA7	SLE	GW	0.92	INT	ND	ND			
289	ENSG00000063241	ISOC2	SLE	GW	0.92	INT	+	ND			
290	ENSG00000197483	ZNF628	SLE	GW	0.79	INT	=	ND			
291	ENSG00000090971	NAT14	SLE	GW	0.91	INT	=	ND			
292	ENSG00000179954	SSC5D	SLE	GW	0.91	INT	ND	ND			
293	ENSG00000179943	FIZ1	SLE	GW	0.78	INT	=	ND			
294	ENSG00000171443	ZNF524	SLE	GW	0.78	INT	-	ND			
295	ENSG00000213015	ZNF580	SLE	GW	0.78	INT	=	ND			

Continued on next page

#	ENSG	Name	Disease	Analysis	COGS	Context	Expr	eRNA	Locus	SNP	p-value
296	ENSG00000171425	ZNF581	SLE	GW	0.78	INT	=	ND			
297	ENSG00000173581	CCDC106	SLE	GW	0.81	INT	ND	ND			
298	ENSG00000175063	UBE2C	CEL	IC	0.54	INT	-	ND			
299	ENSG00000101470	TNNC2	CEL	IC	0.54	INT	ND	ND			
300	ENSG00000124104	SNX21	CEL	IC	0.54	INT	=	ND			
301	ENSG00000198026	ZNF335	CEL	IC	0.92	PROM	=	ND			
302	ENSG00000062598	ELMO2	CEL	IC	0.54	INT	=	ND			
303	ENSG00000159110	IFNAR2	RA	GW	0.84	INT	-	ND	21q22.12	rs2834512	2.1E-08
304	ENSG00000100099	HPS4	SLE	GW	0.66	INT	-	ND			
305	ENSG00000100104	SRRD	SLE	GW	0.66	INT	+	ND			
306	ENSG00000100109	TFIP11	SLE	GW	0.65	INT	+	ND			
307	ENSG00000187045	TMPRSS6	CEL	IC	0.55	INT	ND	=			
308	ENSG00000187045	TMPRSS6	RA	IC	0.64	ACT	ND	ND	22q13.1	rs909685	1.4E-16
309	ENSG00000187045	TMPRSS6	T1D	IC	0.86	INT	ND	ND	22q12.3	rs229533	1.8E-08
310	ENSG00000100385	IL2RB	CEL	IC	0.55	INT	+	=			
311	ENSG00000133466	C1QTNF6	T1D	IC	0.69	PROM	=	ND	22q12.3	rs229533	1.8E-08
312	ENSG00000166897	ELFN2	CEL	IC	0.55	INT	-	=			
313	ENSG00000100060	MFNG	CEL	IC	0.55	INT	-	=			
314	ENSG00000100321	SYNGR1	RA	GW	0.69	PROM	ND	ND	22q13.1	rs909685	1.4E-16

Appendix C

C.1 Relationship between PCA and SVD

Consider a $n \times p$ matrix $\boldsymbol{X}$, where the columns are mean centred. The Principal components of $\boldsymbol{X}$ can be computed by performing eigen-decomposition on the variance-covariance matrix of $\boldsymbol{X}$ such that

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n-1} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T, \tag{C.1}$$

where $\boldsymbol{V}$ is a $p \times n$ matrix where each column represents an eigenvector, that defines a principal component and $\boldsymbol{D}$ is a $n \times p$ diagonal matrix of eigenvalues that reflect the square-root of the variance explained by each component.

Consider the singular value decomposition $\boldsymbol{X}$

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{C.2}$$

where $\boldsymbol{U}$ is the square left singular matrix ($n \times n$), $\boldsymbol{\Sigma}$ is a diagonal matrix of the singular values of $\boldsymbol{X}$ ($n \times n$) and $\boldsymbol{V}$ is the right singular matrix of $\boldsymbol{X}$ ($p \times n$). From this it follows that

$$\begin{aligned}
\boldsymbol{X}^T X &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}, \\
&= \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\right)^T \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \text{Where } \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T.
\end{aligned} \tag{C.3}$$

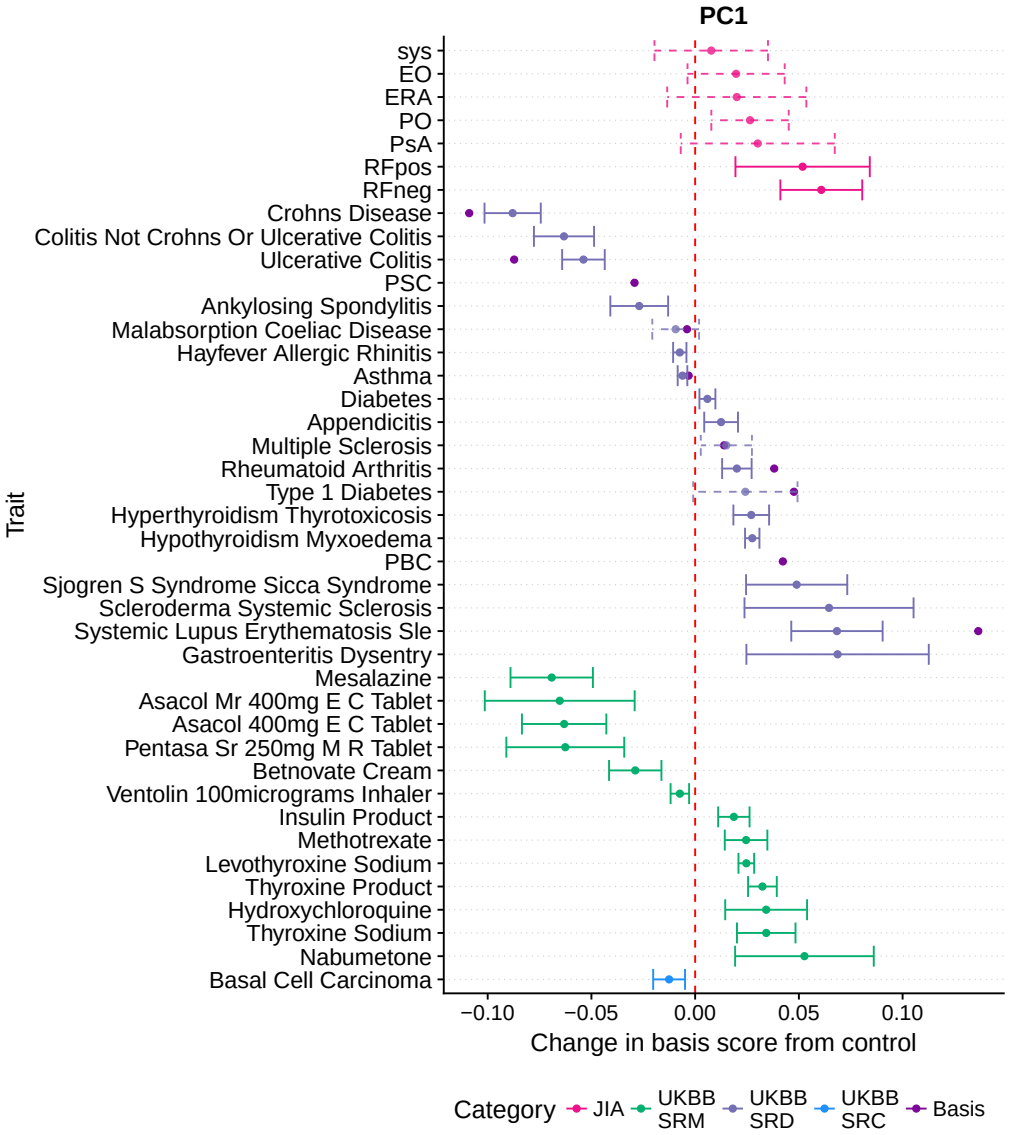
Thus we can relate the eigenvalues of the variance-covariance matrix to the singular values from SVD by

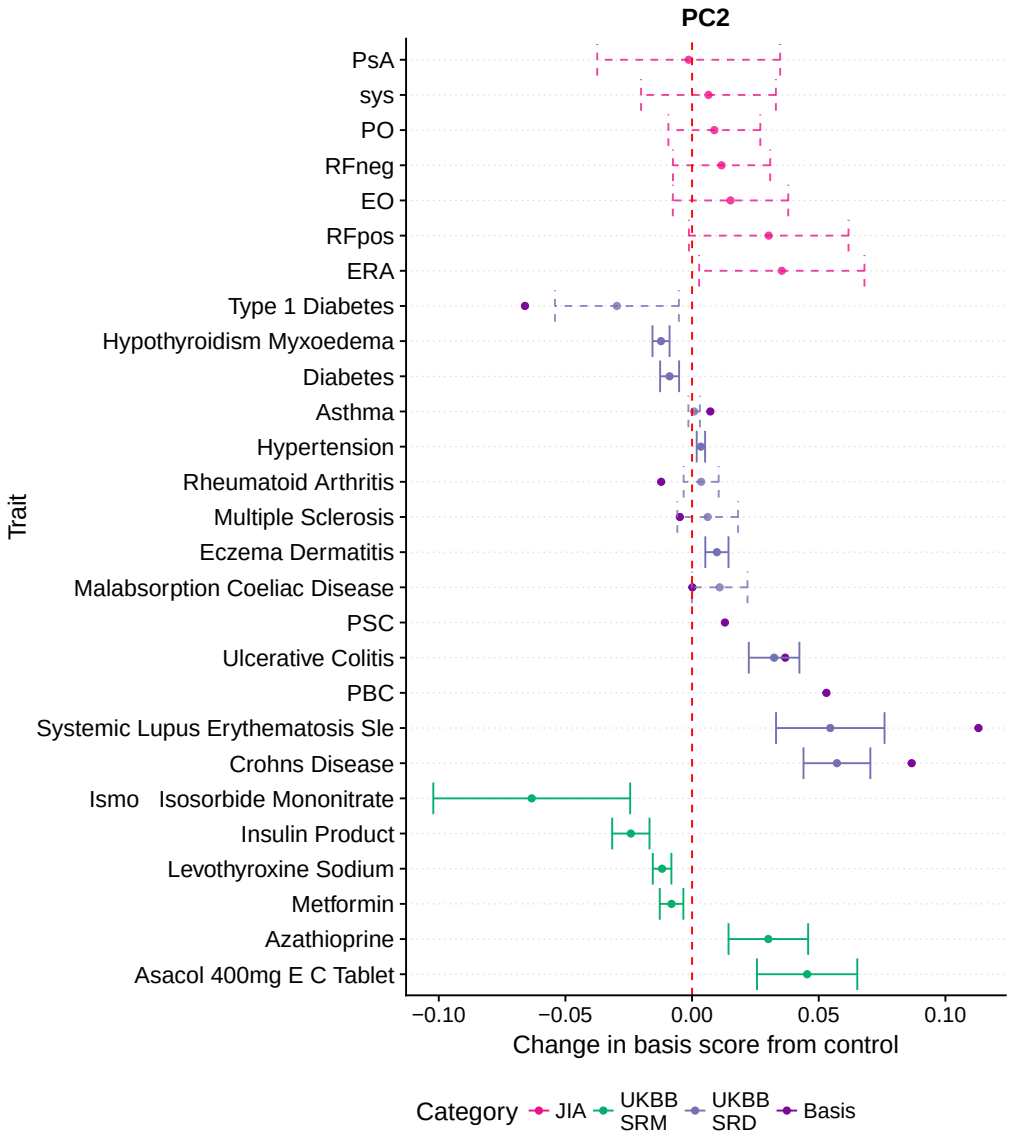
$$\boldsymbol{D} = \frac{\boldsymbol{\Sigma}^2}{n-1}, \tag{C.4}$$

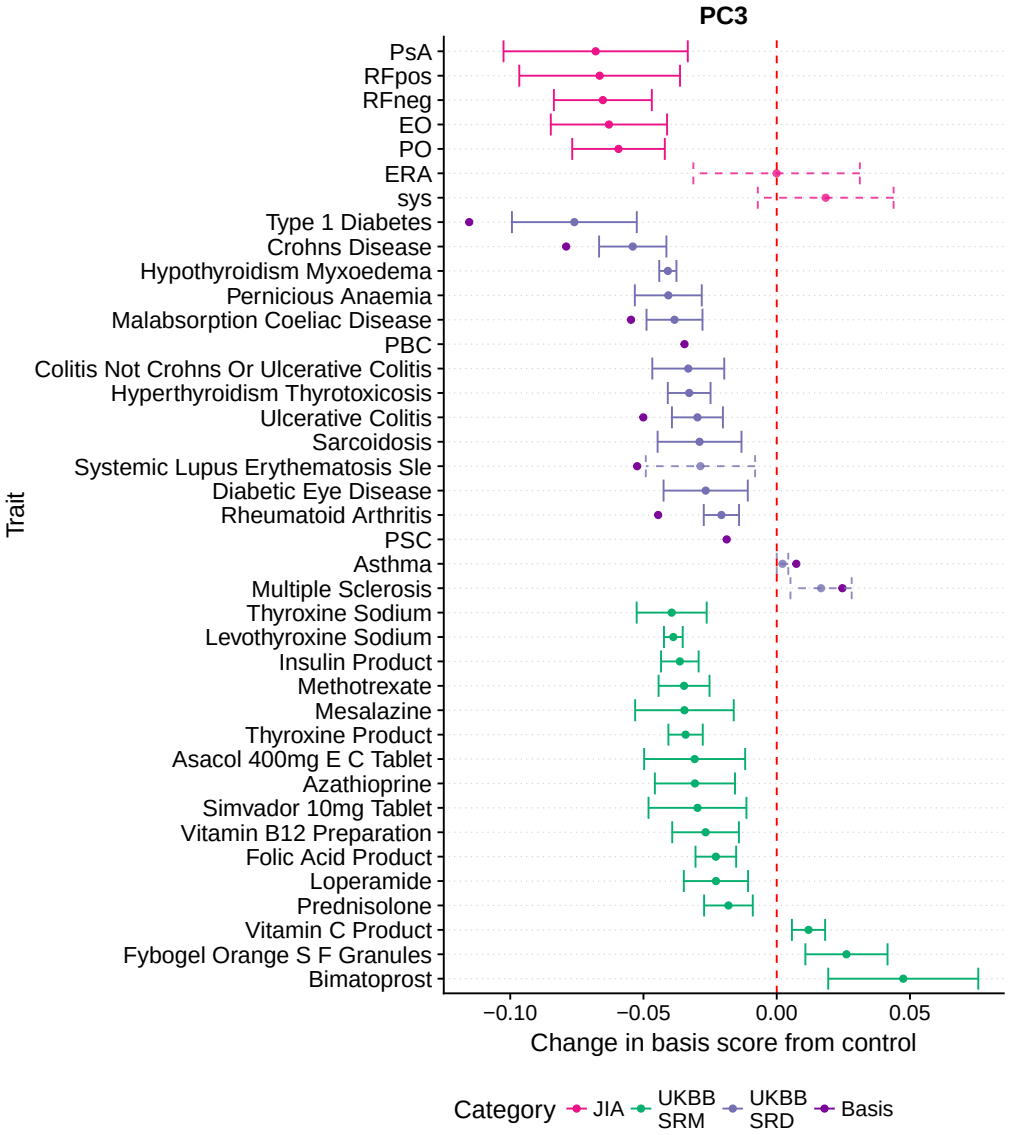
where principal components are given by $\boldsymbol{XV} = \boldsymbol{U\Sigma V}^T\boldsymbol{V} = \boldsymbol{U\Sigma}$ as $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$.

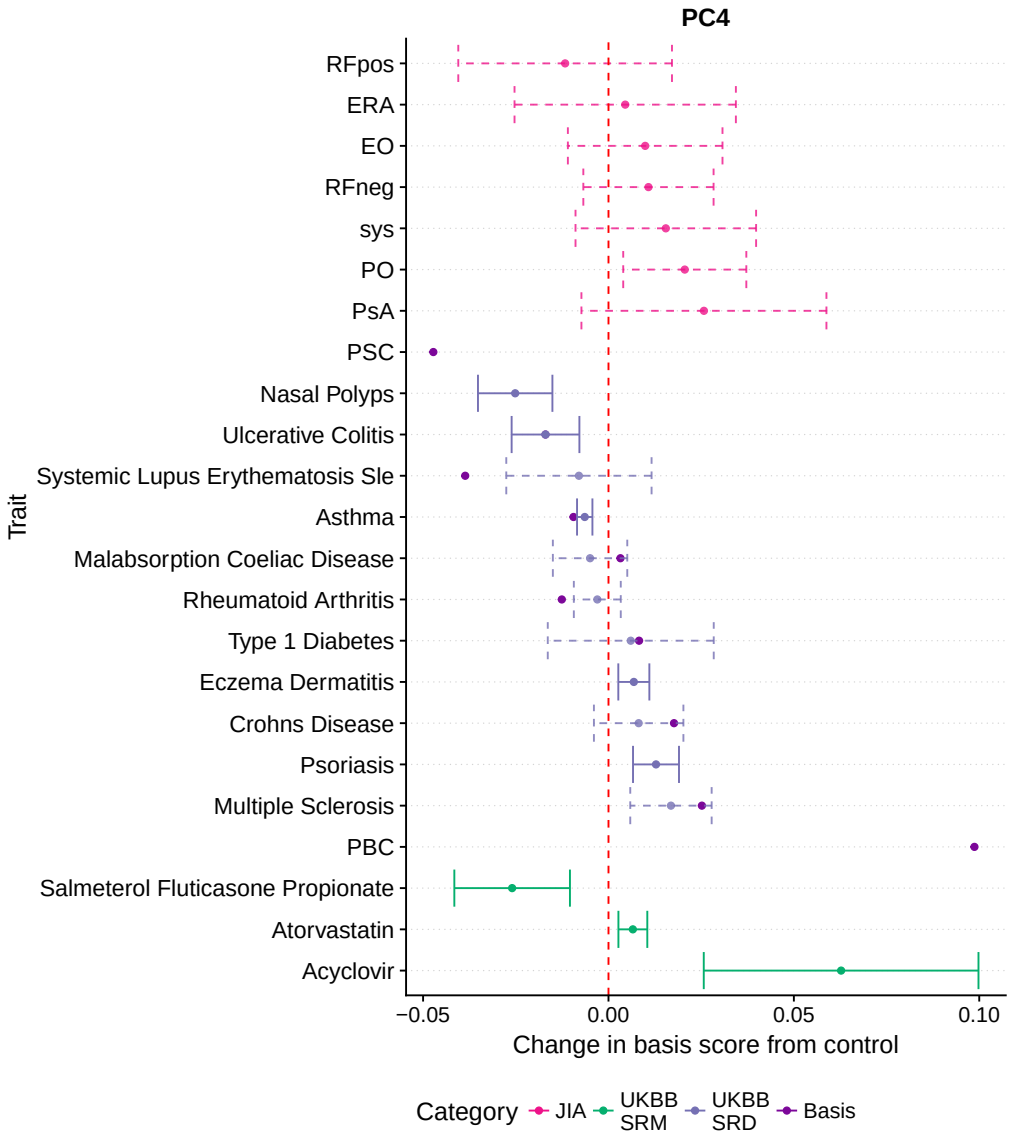
C.2 IMD basis projection forest plots

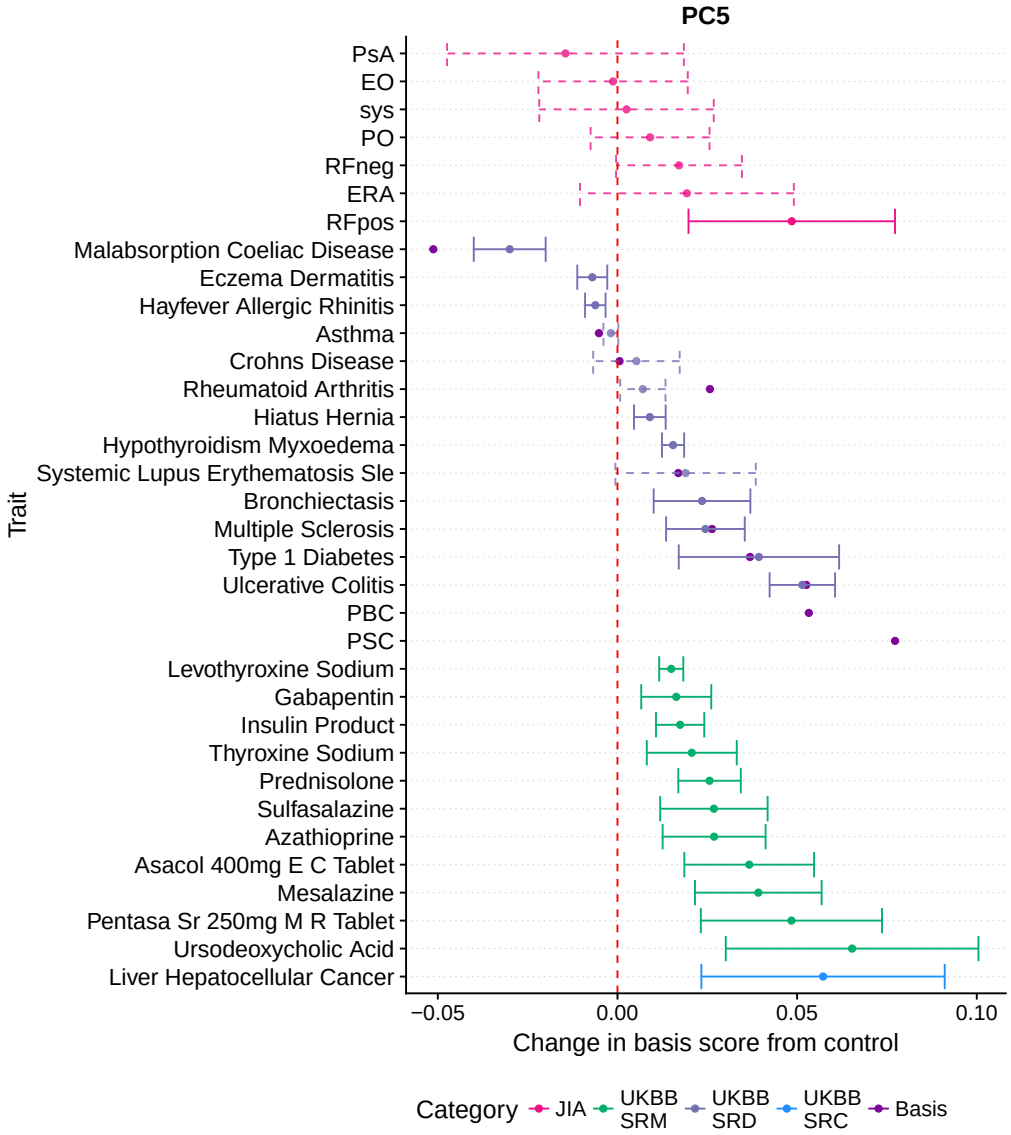
Forest plots showing the context of JIA subtype projections for all PCs. Coloured points indicate the difference in trait scores with synthetic control, marked with a red dotted line. Error bars indicate 95% confidence intervals. For clarity basis trait PC scores (purple) have been merged with overlapping UKBB traits. Whilst non-basis UKBB traits have been filtered to show only those that are significantly different from control at FDR < 5%, other traits are included regardless, these are shown with dashed error bars. UKBB SRD, UKBB SRM and UKB SRC labels in the legend correspond to UK Biobank self-reported disease, medication and cancer categories respectively.

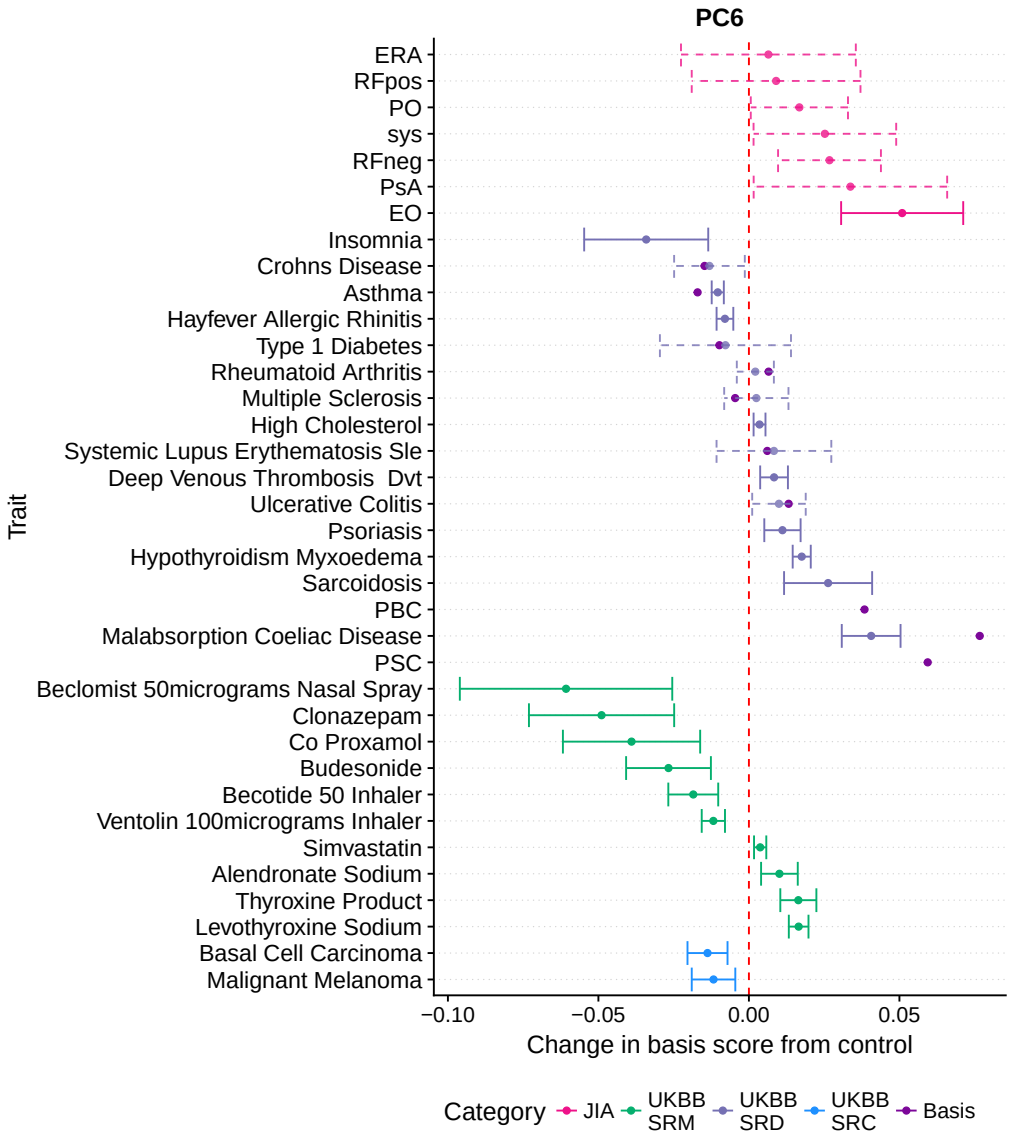

PC1
sys
EO
ERA
PO
PsA
RFpos
RFneg
Crohns Disease
Colitis Not Crohns Or Ulcerative Colitis
Ulcerative Colitis
PSC
Ankylosing Spondylitis
Malabsorption Coeliac Disease
Hayfever Allergic Rhinitis
Asthma
Diabetes
Appendicitis
Multiple Sclerosis
Rheumatoid Arthritis
Type 1 Diabetes
Hyperthyroidism Thyrotoxicosis
Hypothyroidism Myxoedema
PBC
Sjogren S Syndrome Sicca Syndrome
Scleroderma Systemic Sclerosis
Systemic Lupus Erythematosis Sle
Gastroenteritis Dysentry
Mesalazine
Asacol Mr 400mg E C Tablet
Asacol 400mg E C Tablet
Pentasa Sr 250mg M R Tablet
Betnovate Cream
Ventolin 100micrograms Inhaler
Insulin Product
Methotrexate
Levothyroxine Sodium
Thyroxine Product
Hydroxychloroquine
Thyroxine Sodium
Nabumetone
Basal Cell Carcinoma
Trait
−0.10
−0.05
0.00
0.05
0.10
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
UKBB SRC
Basis

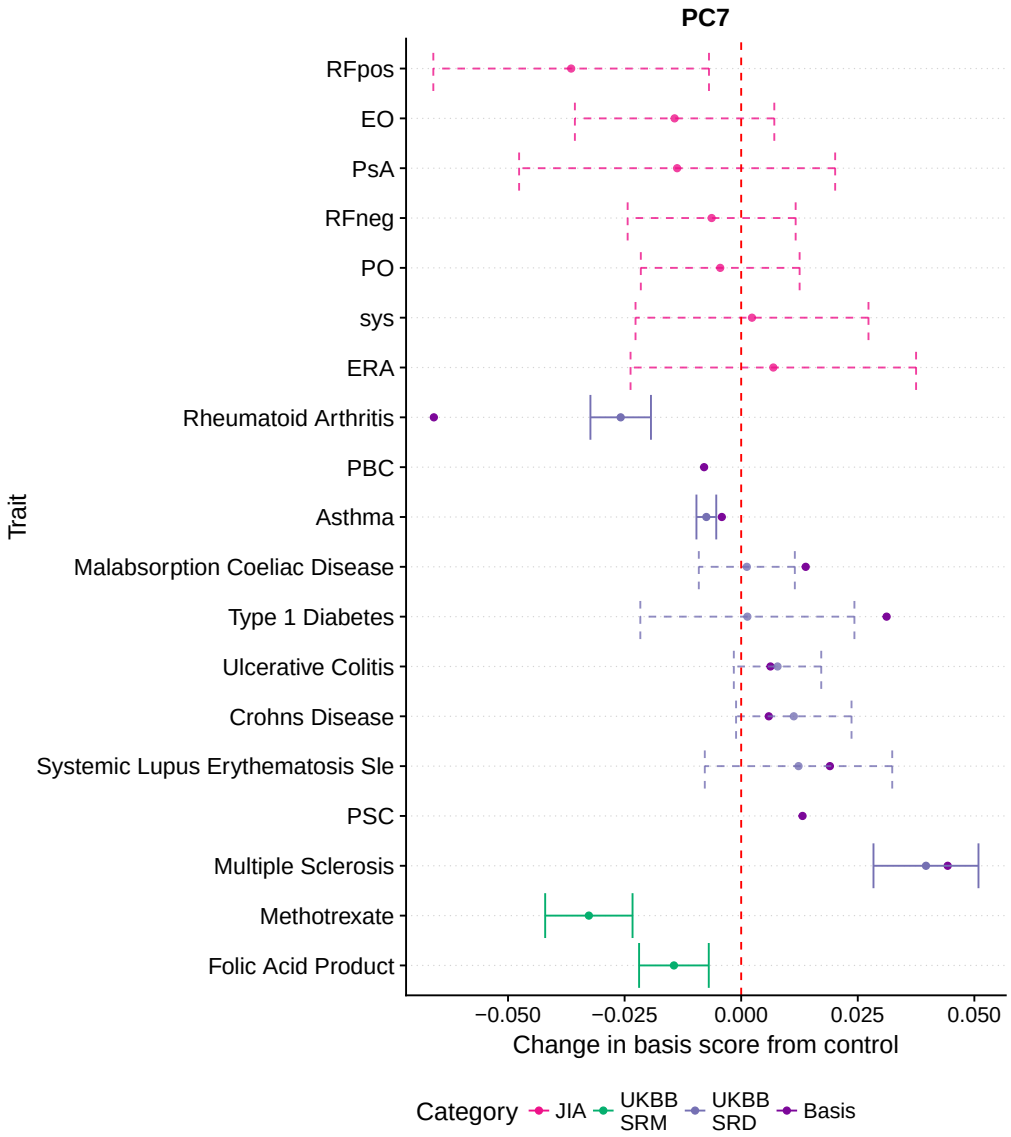

PC2
PsA
sys
PO
RFneg
EO
RFpos
ERA
Type 1 Diabetes
Hypothyroidism Myxoedema
Diabetes
Asthma
Hypertension
Rheumatoid Arthritis
Multiple Sclerosis
Eczema Dermatitis
Malabsorption Coeliac Disease
PSC
Ulcerative Colitis
PBC
Systemic Lupus Erythematosis Sle
Crohns Disease
Ismo Isosorbide Mononitrate
Insulin Product
Levothyroxine Sodium
Metformin
Azathioprine
Asacol 400mg E C Tablet
Trait
−0.10
−0.05
0.00
0.05
0.10
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis

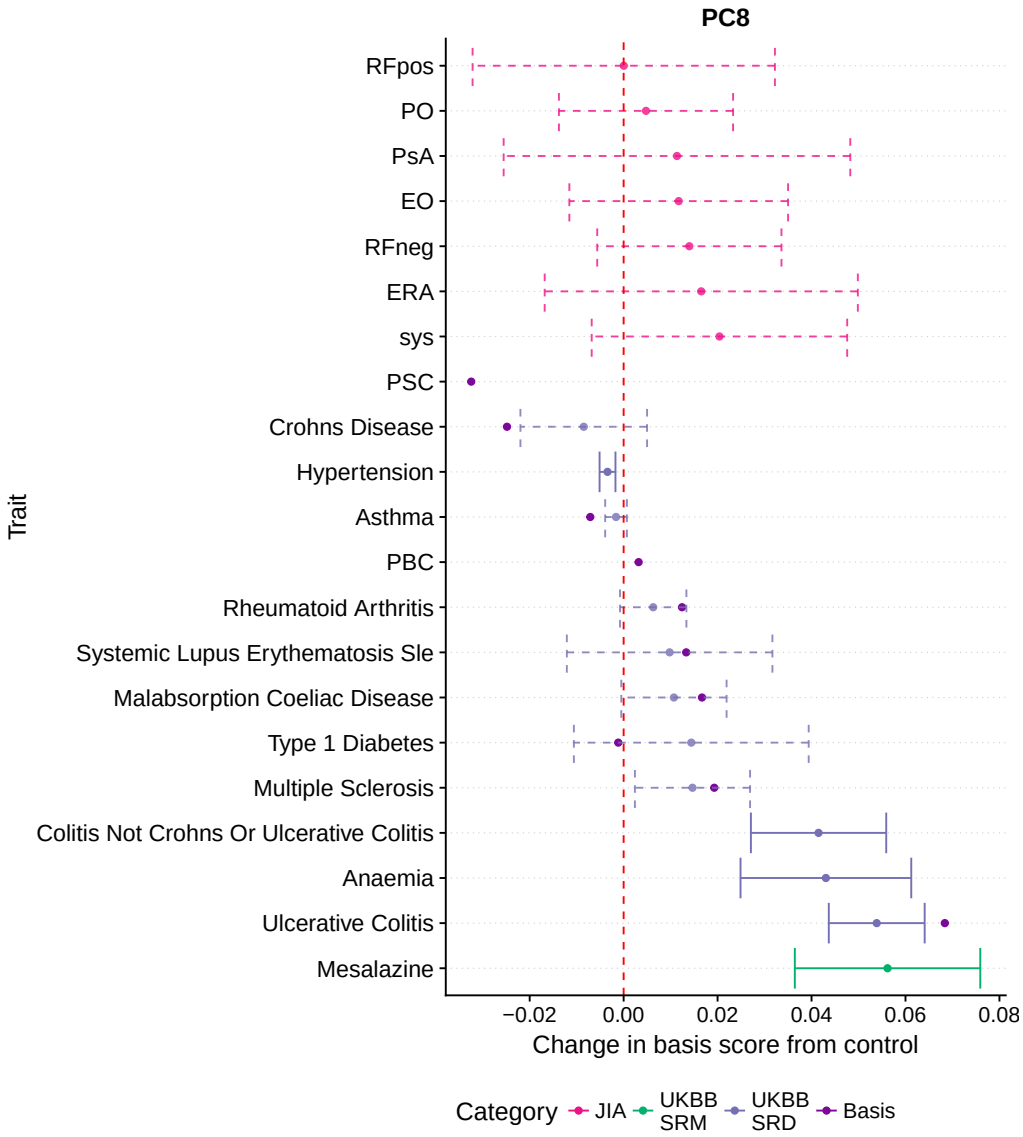

PC3
PsA
RFpos
RFneg
EO
PO
ERA
sys
Type 1 Diabetes
Crohns Disease
Hypothyroidism Myxoedema
Pernicious Anaemia
Malabsorption Coeliac Disease
PBC
Colitis Not Crohns Or Ulcerative Colitis
Hyperthyroidism Thyrotoxicosis
Ulcerative Colitis
Sarcoidosis
Systemic Lupus Erythematosis Sle
Diabetic Eye Disease
Rheumatoid Arthritis
PSC
Asthma
Multiple Sclerosis
Thyroxine Sodium
Levothyroxine Sodium
Insulin Product
Methotrexate
Mesalazine
Thyroxine Product
Asacol 400mg E C Tablet
Azathioprine
Simvador 10mg Tablet
Vitamin B12 Preparation
Folic Acid Product
Loperamide
Prednisolone
Vitamin C Product
Fybogel Orange S F Granules
Bimatoprost
Trait
−0.10
−0.05
0.00
0.05
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis

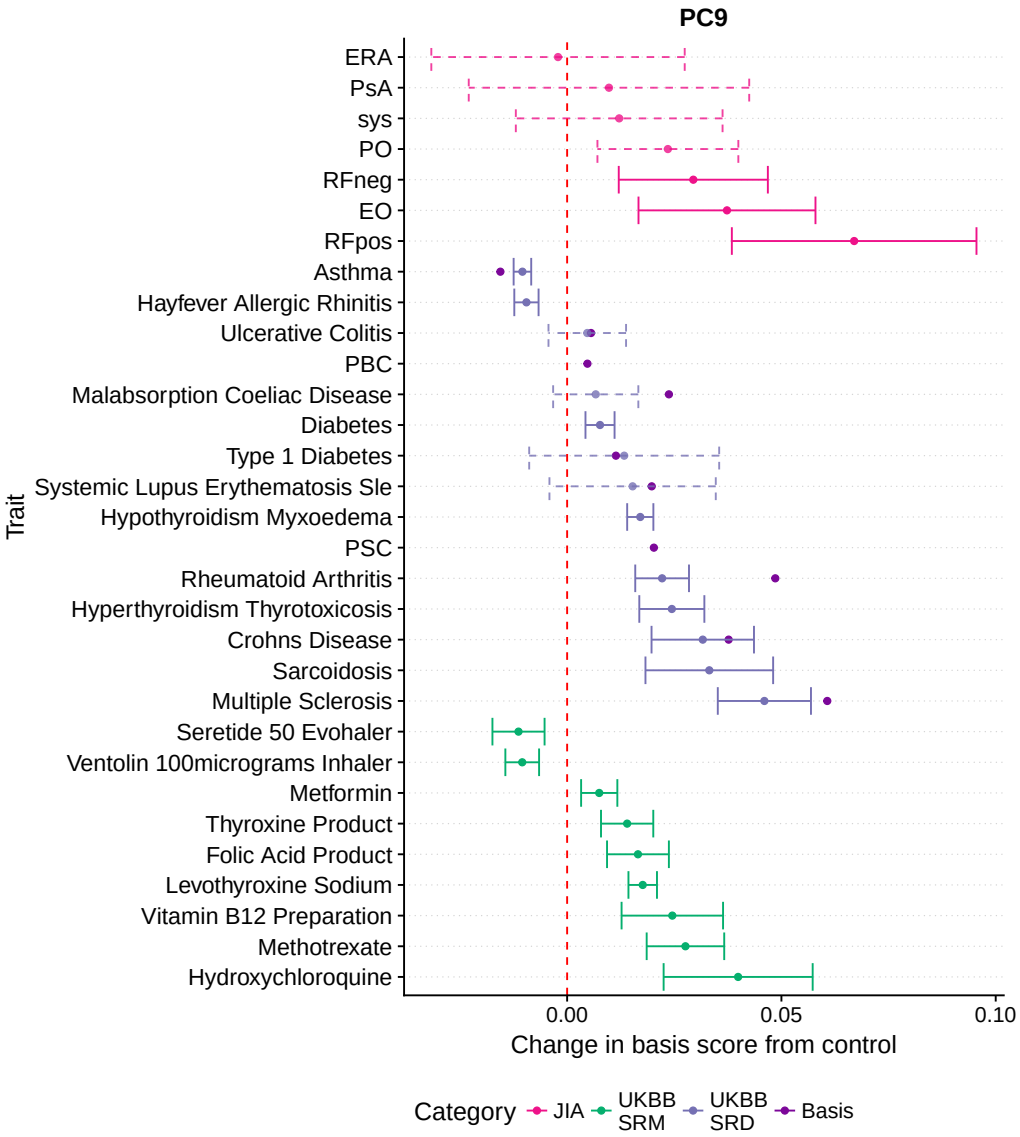

PC4
RFpos
ERA
EO
RFneg
sys
PO
PsA
PSC
Nasal Polyps
Ulcerative Colitis
Systemic Lupus Erythematosis Sle
Asthma
Malabsorption Coeliac Disease
Rheumatoid Arthritis
Type 1 Diabetes
Eczema Dermatitis
Crohns Disease
Psoriasis
Multiple Sclerosis
PBC
Salmeterol Fluticasone Propionate
Atorvastatin
Acyclovir
Trait
−0.05
0.00
0.05
0.10
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis

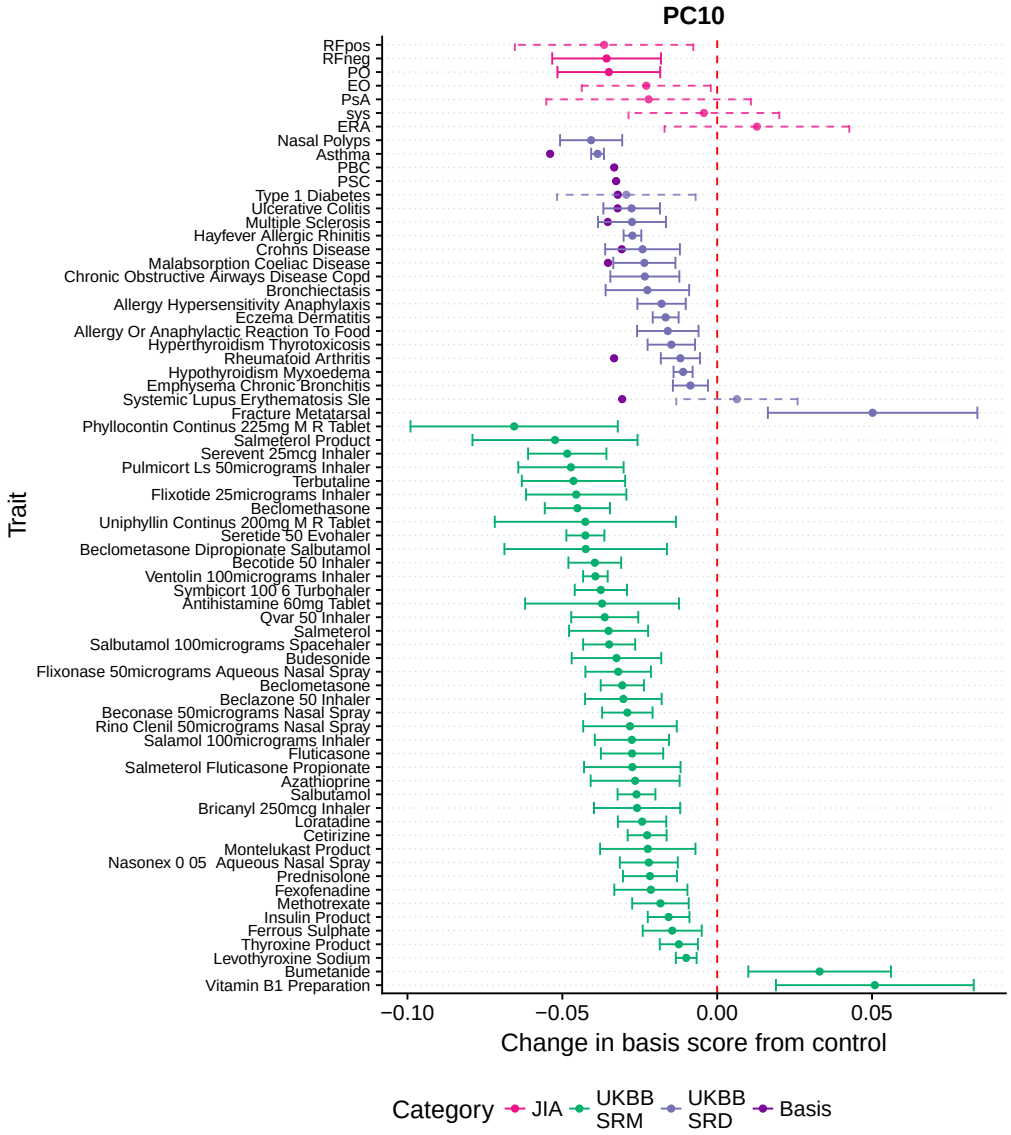

PC5
PsA
EO
sys
PO
RFneg
ERA
RFpos
Malabsorption Coeliac Disease
Eczema Dermatitis
Hayfever Allergic Rhinitis
Asthma
Crohns Disease
Rheumatoid Arthritis
Hiatus Hernia
Hypothyroidism Myxoedema
Systemic Lupus Erythematosis Sle
Bronchiectasis
Multiple Sclerosis
Type 1 Diabetes
Ulcerative Colitis
PBC
PSC
Levothyroxine Sodium
Gabapentin
Insulin Product
Thyroxine Sodium
Prednisolone
Sulfasalazine
Azathioprine
Asacol 400mg E C Tablet
Mesalazine
Pentasa Sr 250mg M R Tablet
Ursodeoxycholic Acid
Liver Hepatocellular Cancer
Trait
−0.05
0.00
0.05
0.10
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
UKBB SRC
Basis


PC6
ERA
RFpos
PO
sys
RFneg
PsA
EO
Insomnia
Crohns Disease
Asthma
Hayfever Allergic Rhinitis
Type 1 Diabetes
Rheumatoid Arthritis
Multiple Sclerosis
High Cholesterol
Systemic Lupus Erythematosis Sle
Deep Venous Thrombosis Dvt
Ulcerative Colitis
Psoriasis
Hypothyroidism Myxoedema
Sarcoidosis
PBC
Malabsorption Coeliac Disease
PSC
Beclomist 50micrograms Nasal Spray
Clonazepam
Co Proxamol
Budesonide
Becotide 50 Inhaler
Ventolin 100micrograms Inhaler
Simvastatin
Alendronate Sodium
Thyroxine Product
Levothyroxine Sodium
Basal Cell Carcinoma
Malignant Melanoma
Trait
−0.10
−0.05
0.00
0.05
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
UKBB SRC
Basis


PC7
RFpos
EO
PsA
RFneg
PO
sys
ERA
Rheumatoid Arthritis
PBC
Asthma
Malabsorption Coeliac Disease
Type 1 Diabetes
Ulcerative Colitis
Crohns Disease
Systemic Lupus Erythematosis Sle
PSC
Multiple Sclerosis
Methotrexate
Folic Acid Product
Trait
−0.050
−0.025
0.000
0.025
0.050
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis


PC8
RFpos
PO
PsA
EO
RFneg
ERA
sys
PSC
Crohns Disease
Hypertension
Asthma
PBC
Rheumatoid Arthritis
Systemic Lupus Erythematosis Sle
Malabsorption Coeliac Disease
Type 1 Diabetes
Multiple Sclerosis
Colitis Not Crohns Or Ulcerative Colitis
Anaemia
Ulcerative Colitis
Mesalazine
Trait
−0.02
0.00
0.02
0.04
0.06
0.08
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis


PC9
ERA
PsA
sys
PO
RFneg
EO
RFpos
Asthma
Hayfever Allergic Rhinitis
Ulcerative Colitis
PBC
Malabsorption Coeliac Disease
Diabetes
Type 1 Diabetes
Systemic Lupus Erythematosis Sle
Hypothyroidism Myxoedema
PSC
Rheumatoid Arthritis
Hyperthyroidism Thyrotoxicosis
Crohns Disease
Sarcoidosis
Multiple Sclerosis
Seretide 50 Evohaler
Ventolin 100micrograms Inhaler
Metformin
Thyroxine Product
Folic Acid Product
Levothyroxine Sodium
Vitamin B12 Preparation
Methotrexate
Hydroxychloroquine
Trait
0.00
0.05
0.10
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis


PC10
Trait
RFpos
RFneg
PO
EO
PsA
sys
ERA
Nasal Polyps
Asthma
PBC
PSC
Type 1 Diabetes
Ulcerative Colitis
Multiple Sclerosis
Hayfever Allergic Rhinitis
Crohns Disease
Malabsorption Coeliac Disease
Chronic Obstructive Airways Disease Copd
Bronchiectasis
Allergy Hypersensitivity Anaphylaxis
Eczema Dermatitis
Allergy Or Anaphylactic Reaction To Food
Hyperthyroidism Thyrotoxicosis
Rheumatoid Arthritis
Hypothyroidism Myxoedema
Emphysema Chronic Bronchitis
Systemic Lupus Erythematosis Sle
Fracture Metatarsal
Phyllocontin Continus 225mg M R Tablet
Salmeterol Product
Serevent 25mcg Inhaler
Pulmicort Ls 50micrograms Inhaler
Terbutaline
Flixotide 25micrograms Inhaler
Beclomethasone
Uniphyllin Continus 200mg M R Tablet
Seretide 50 Evohaler
Beclometasone Dipropionate Salbutamol
Becotide 50 Inhaler
Ventolin 100micrograms Inhaler
Symbicort 100 6 Turbohaler
Antihistamine 60mg Tablet
Qvar 50 Inhaler
Salmeterol
Salbutamol 100micrograms Spacehaler
Budesonide
Flixonase 50micrograms Aqueous Nasal Spray
Beclometasone
Beclazone 50 Inhaler
Beconase 50micrograms Nasal Spray
Rino Clenil 50micrograms Nasal Spray
Salamol 100micrograms Inhaler
Fluticasone
Salmeterol Fluticasone Propionate
Azathioprine
Salbutamol
Bricanyl 250mcg Inhaler
Loratadine
Cetirizine
Montelukast Product
Nasonex 0 05 Aqueous Nasal Spray
Prednisolone
Fexofenadine
Methotrexate
Insulin Product
Ferrous Sulphate
Thyroxine Product
Levothyroxine Sodium
Bumetanide
Vitamin B1 Preparation
−0.10
−0.05
0.00
0.05
Change in basis score from control
Category
JIA
UKBB SRM
UKBB SRD
Basis